Ph.D. DISSERTATION

3D Reconstruction of Multiple Objects from Dynamic Scenes and Learning Based Depth Super Resolution

동적 장면으로부터의 다중 물체 3차원 복원 기법 및 학습 기반의 깊이 초해상도 기법

BY

YOUNG MIN SHIN

February 2014

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

3D Reconstruction of Multiple Objects from Dynamic Scenes and Learning Based Depth Super Resolution

동적 장면으로부터의 다중 물체 3차원 복원 및
학습 기반의 깊이 초해상도 기법

지도교수 이 경 무
이 논문을 공학박사 학위논문으로 제출함
2014 년 2 월

서울대학교 대학원
전기컴퓨터공학부
신 영 민

신영민의 공학박사 학위논문을 인준함
2014 년 2 월

위 원 장 : 李 商 郎 Sang-Mk Lee
부위원장 : 李 恩 式
위 원 : 李 相 旭 Sanguk
위 원 : 朴 仁 圭
위 원 : 林 鍾 宇

Abstract

In this dissertation, a framework for reconstructing 3-dimensional shape of the multiple objects and the method for enhancing the resolution of 3-dimensional models, especially human face, are proposed. Conventional 3D reconstruction from multiple views is applicable to static scenes, in which the configuration of objects is fixed while the images are taken. In the proposed framework, the main goal is to reconstruct the 3D models of multiple objects in a more general setting where the configuration of the objects varies among views. This problem is solved by object-centered decomposition of the dynamic scenes using unsupervised co-recognition approach. Unlike conventional motion segmentation algorithms that require small motion assumption between consecutive views, co-recognition method provides reliable accurate correspondences of a same object among unordered and wide-baseline views. In order to segment each object region, the 3D sparse points obtained from the structure-from-motion are utilized. These points are relative reliable since both their geometric relation and photometric consistency are considered simultaneously to generate these 3D sparse points. The sparse points serve as automatic seed points for a seeded-segmentation algorithm, which makes the interactive segmentation work in non-interactive way. Experiments on various real challenging image sequences demonstrate the effectiveness of the proposed approach, especially in the presence

of abrupt independent motions of objects.

Obtaining high-density 3D model is also an important issue. Since the multi-view images used to reconstruct 3D model or the 3D imaging hardware such as the time-of-flight cameras or the laser scanners have their own natural upper limit of resolution, super-resolution method is required to increase the resolution of 3D data. This dissertation presents an algorithm to super-resolve the single human face model represented in 3D point cloud. The point cloud data is considered as an object-centered 3D data representation compared to the camera-centered depth images. While many researches are done for the super-resolution of intensity images and there exist some prior works on the depth image data, this is the first attempt to super-resolve the single set of 3D point cloud data without additional intensity or depth image observation of the object. This problem is solved by querying the previously learned database which contains corresponding high resolution 3D data associated with the low resolution data. The Markov Random Field(MRF) model is constructed on the 3D points, and the proper energy function is formulated as a multi-class labeling problem on the MRF. Experimental results show that the proposed method solves the super-resolution problem with high accuracy.

Key words: Computer Vision, 3D Reconstruction, Dynamic Scenes, Co-recognition, Multiple Objects, Super-resolution, Point Cloud

Student number: 2009-30195

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 3D Computer Vision

Since human has 3 dimensional (3D) visual perception system, 3D computer vision problems such as stereo matching, multi view geometry, dense 3D reconstruction, and augmented reality, are fundamental research topics. The images are the first and foremost important observation source. As the images are produced by the projection of real 3D world to the 2D image plane, understanding 3D world from the 2D image data is quite difficult and challenging task. Many researchers dedicated to raise the 3D computer models from the images of existing objects, and their effort made substantial progress in 3D reconstruction. However, in the past few decades the majority of researches on the 3D computer vision focused on reconstructing the objects in the controlled settings.

Therefore, this dissertation addresses the studies on the methods to solve 3D computer vision problem in a various environments. Firstly, a novel integrated sys-

tem is proposed to obtain 3D models from the less controlled dynamic scenes captured by a monocular camera. Conventionally, the general target scenes are considered globally static to avoid feature mismatching. Although the motion segmentation is addressed to deal with the motions in the target scene, they were unable to handle discontinuities or abrupt motion. The proposed system adopts object recognition based feature exploration technique to 3D computer vision framework to solve the multiple motion problem. Secondly, a new method for enhancing 3D point cloud data is presented for human face data. Since the point cloud is the viewpoint-invariant representation of object shape, it has many advantages over the viewpoint-oriented depth images. However the point cloud causes many difficulties in processing the patches due to its free-form representation. This dissertation describes a super resolution method which utilizes the exemplar database constructed from nonparametric training data.

1.2 Dissertation Goal and Contribution

The goal of the dissertation is to apply object recognition technique to build a novel object-centered 3D reconstruction system which is applicable to more general scenes containing multiple dynamic objects, and to study a new super resolution method to enhance the resolution of built 3D model, especially for human faces. The contribution of the dissertation is summarized as follows:

- A framework for reconstructing 3-dimensional shape of the multiple objects is proposed. Conventional 3D reconstruction from multi-view images is only applicable to static scenes. However, in the proposed frame-

work, 3D models of multiple object is obtained in a more general setting where the configuration of the objects varies among views. The image exploration method expands and clusters initial feature matches up to the object-level correspondences. The information required for refining segmentation mask is also obtained from the sparse 3D reconstruction. The proposed method overcomes the troublesome abrupt motion issue of feature tracking based methods, and shows 3D reconstruction results from both video clips and independently captured images.

- A novel 3D super resolution method for 3D point cloud data is presented. There are only a few studies on the super resolution of the 3D data. The prior works on the 3D super resolution method tried to enhance the resolution of the depth images, which are actually camera-centered data. In the proposed method, a point cloud data which contains object-centered 3D shape information is super-resolved in a non-parametric exemplar based approach. To utilize local approach in the 3D point cloud, efficient methods treating the 3D local patches are proposed. In this study, the experiments are performed on the human face data.

1.3 Organization of Dissertation

The remainder of this dissertation is structured as follows. In chapter 2, the motives and fundamentals to understand the proposed method is briefly introduced with prominent prior works.

The main body of this dissertation consists of two related line of works, one for

multiple view geometry reconstruction framework applied for several independent objects in dynamic scenes, and one for super resolution of 3D model by utilizing exemplar based up-sampling, concentrating on the human face, which is the prominent interest of human visual perception system. Each work constitutes each chapter of the dissertation.

In chapter 3, a 3D reconstruction system based on the object recognition, which solves the problem of estimating object number and feature matching is described [3]. When a scene with multiple dynamic objects is reconstructed, each independent motion of object raises various challenges. In the conventional setting, the target scene should maintain static during the image acquisition to meet the global geometric consistency, which is essential for 3D reconstruction method. The static constraint, which is an implicit assumption for the most of conventional 3D reconstruction framework, is not satisfied in the dynamic scenes. In the proposed method, the object recognition based approach decomposes the target scene into meaningful object regions with object level correspondences. The proposed algorithm achieves the object-centered 3D reconstruction from various input images and videos.

Chapter 4 presents a novel 3D super resolution method applied for 3D face point cloud. When a visual signal is properly down-sampled, we lose the information about the detail of the target. In terms of the frequency analysis, the high-frequency component which includes the detail is missing in the down-sampled, or low-resolution signal. Estimating the high-resolution data from the low-resolution data is one of the *inverse problems* in computer vision. To fill in the information gap, prior knowledge or some external information is required. The proposed method is based on the

exemplar database built upon many off-line observation data. The proposed non-parametric approach effectively restored the detail from the severely impaired 3D point cloud data.

Finally, we conclude the dissertation in chapter 5.

Chapter 2

Background

Since the 3D visual data gives novel experience to people in comparison with 2D images, the 3D data is becoming more important. The 3D information can be utilized in practical applications such as free viewpoint video [4], or image relighting [5] etc. Moreover, the propagation of low-priced direct 3D acquisition devices such as Kinect, time-of-fight(TOF) depth cameras or the 3D scanners accelerate this tendency.

Among the advance of 3D sensors, image is still playing the key role to obtain 3D information. As the cell phones bring ubiquitous cameras and the Internet is used to share digital data, the digital photos are now flooding the world. Many archives contain image data from historic to personal records, which are actually 2D projection of 3D world.

In this section, some prominent prior works proposed for the multi-body motion problem and the image super resolution, which are considered as basis for the multi-object reconstruction and 3D super-resolution.

2.1 Motion Segmentation

While the 3D geometry and reconstruction has been research topic for a long time, only a few have focused directly on the dynamic scenes. The topic of motion segmentation has to be addressed when dealing with the multi-body dynamic scene problem. Model selection and subspace separation are common concepts in this area. Wang and Adelson [6] presented the idea of assigning pixels to overlapping layers, where each layer's motion is described by a smooth flow field. This has been referred to as a layered approach, which explains pixel motions as a parametric motion model featuring several layers. Such representation has been adopted by a number of algorithms. They have often used expectation-maximization (EM) [7] [8] or graph cuts [2] [9] to minimize energy functions.

Some works [10] [11] [12] approached the multiple motion segmentation problem as mathematical multi-body factorization. They are based on the subspace constraints that the trajectories of points on the independent rigid objects are from independent subspaces. The factorization method is primarily focused on the algebraic explanation to the problem; the authors assumed that the feature-tracking issue has been solved, in fact, this is an unrealistic assumption. Most studies on motion segmentation tend to concentrate on theoretic aspects and have derived solutions from simple data with restricted environments.

In response to the large disparity discrete motion problem, Wills et al. [2] presented a method that combined the layer-based approach concept and feature-based motion estimation. First, the initial correspondences matched by comparing the descriptor vectors of interest points are computed. Established initial matches are

perturbed to check correctness and boost the inlier matches. They then used a RANSAC-based procedure to detect and partition the motion fields of the frames. Finally, an approximate graph cut method is applied to assign pixels densely to each motion field. Their proposed approach demonstrated the ability to handle large inter-frame motion, which is the limitation of the optical flow-based feature tracking methods. The work of Wills et al. is similar to our method in that it can handle large abrupt motions and boost true inlier interest point matches. However, their work was based on a strong assumption that objects are matched by a single planar homography between images. This assumption does not hold for most general 3D objects, particularly when images are taken under a wide baseline setting.

2.2 Image Super Resolution

In connection with the super resolution, this technique has been arisen from the intensity image up-sampling method. To enlarge the size of the image, some simple interpolation methods like linear, cubic, nearest neighbor interpolation scheme are the simplest ones. However, these kinds of methods do not really improve the quality of the image in terms of the amount of information. They only generate new intermediate pixels by duplicating the information of the nearby pixels or by calculating the average value between them.

Traditionally, the super resolution problem focuses on aggregating multiple low-resolution observations of the scene and produces a high-resolution output image. The algorithms in this category of methods are called as *multi-frame* approach [13, 14, 15]. In the multi-frame based approach, we are given a number of low-

resolution images differing in geometric transformations, lighting (photometric) variations, camera blur (point-spread function), image quantization and noise. To utilize the information of the true scene partly pervaded in the multiple low-resolution images, the sub-pixel accuracy registration or blur estimation stage is required. The main theory of the multi-frame based approach is *back projection*. A set of observed images are registered together, and each pixel in the low resolution image projects their intensity values to the high-resolution latent image. Since there are many redundant observation, the pixels in high resolution image combines information received from many different input pixels.

Although the multi-frame approach achieved significant progress in improving the image quality and is based on the theoretically sound mathematics, it cannot be applied when multiple frames are not available or the target scene is not static. Thus, the another category of the algorithms for super-resolving the single low resolution image without multiple observation, known as *single-frame* approach are proposed. Freeman et al. [16, 17] solved the problem utilizing the pre-learned database of images. They modeled the problem as a discrete labeling optimization on the MRF structure. Small local patches are extracted from the low resolution input image, and each local patch form each node of the MRF. For each local patch, some high resolution patches in the database are selected with the similarity to the corresponding local patch, and the regularization term tries to promote the neighboring high resolution patches to have smooth boundaries. Since the high frequency information, which is missing the the low resolution input image, is transferred from the external exemplar patches, it is also called an exemplar based approach.

Glasner et al. [18] also proposed similar exemplar based super resolution method, but their approach was exploiting the same image. Their approach is based on the idea that similar visual patterns can be found in the same image instead of the external images. They showed the super resolution results of only single image, that contains many repetitive pattern, without external image database. Later, Zontak and Irani [19] presented detailed quantification of Glasner's idea. Their study revealed that an extracted patch is more likely to occur again in the same image, enjoying obvious advantage over external images. According to their paper, it is *almost sure* that a patch recurs again in the same image while many external images have to be searched to find the same exemplar.

Yang et al. [20] applied the sparse coding technique to super resolution. The image patches are represented as sparse linear combination of over-complete dictionary. In their method, the low-resolution input patch is expressed as its coefficients of the dictionary, and the corresponding high-resolution patch is reconstructed by utilizing the coefficients representation.

Chapter 3

Multi-Object Reconstruction from Dynamic Scenes

3.1 Introduction

Multiple view geometry is an important research area in computer vision. Traditionally, multiple view-based three-dimensional (3D) reconstruction systems are restricted to static scenes, in which the scene configuration is unchanged as the images are taken in different views. In this paper, we aim to develop a reconstruction system that is capable of building 3D models of multiple rigid objects in dynamic scenes where each object and camera are moving freely.

Suppose there are multiple target objects appearing simultaneously in the input images, and they are settled in different configurations in each scene. For example, several independent objects are captured in the multiple images, and the position and pose of these objects vary from image to image. In such settings, the recon-

struction problem from the dynamic scene raises various challenges. Since there are arbitrary motions of multiple target objects in the scene, the multi-view constraint of the *whole* scene is not satisfied. Given that the 3D reconstruction applied to the whole images is not established due to the lack of global geometric consistency, the geometric conflict among the target objects has to be resolved prior to reconstruction. Thus, a natural way is to separate each object from the others and reconstruct each one independently. Once identifying and segmenting corresponding regions of same object among images and establishing feature correspondences among the regions, the multi-body problem can be easily resolved by finding relative camera poses for each object region individually.

To solve the multi-body problem, we approach from the object recognition viewpoint. If we recognize each independently moving object in the image sequences, the solution gives us answers to the following concerns: the number of the objects, the membership of the objects, and their correspondences. These data serve as clues in reconstructing the structure of individual objects in the dynamic scene environment [21]. In the present work, we apply object recognition technique to decompose the dynamic scene into coherent regions that correspond to separated objects distinguished by color or texture characteristics.

The key contributions of our approach are summarized as follows:

1. We designed a 3D reconstruction system for multiple objects in dynamic scenes. It is the first solution to the 3D reconstruction problem of dynamic abrupt scenes with arbitrary view points. The solution to such problem has never been proposed in previous works. Our method is general in that it does not have any

(a) Input images contain two house objects with arbitrary independent poses.

(b) 3D reconstruction results of both *house* objects.

Figure 3.1: An example of multiple object reconstruction from a dynamic scene. Despite the arbitrary positions and poses of the objects in the images, each object is separated and reconstructed to the individual 3D model.

restriction on the number of objects or the ordering of sequences, yet yielding compelling results.

2. Extended co-recognition technique has been utilized for identifying objects in multiple images. Co-recognition avoids feature tracking issues, one of the most troublesome problems in this area, when motion segmentation is applied to real images. Without feature tracking, our method can handle the general settings of objects between images.

3. We have refined the patch-level object boundary obtained from object recognition algorithm up to the pixel-level precision. We applied an interactive image segmentation method in a non-interactive way by providing seed points automatically. The 3D structure of the target object is utilized to get reliable object segment.

The structure of this chapter is as follows: in Section 2, we start with a brief review of related works; Section 3 describes our approach and the detail of the proposed system; we present the experimental results on various data in Section 4; and the paper concludes in Section 5.

3.2 Related Work

Only few studies have focused directly on this subject, and the works of Rothganger et al. [1] and Ozden et al. [22] are the closest ones to the proposed system. Rothganger et al. presented a 3D structure representation using a collection of small planar patches combined with their normalized local appearance description. They segmented a scene into rigid components and constructed 3D models of the components up to the local affine patch level. The major objects are distinguished one

after another using the RANSAC (RANdom SAmple Consensus) procedure on the feature tracks [23]. They applied their representation in detecting moving objects in video sequences. The work of Ozden et al. [22], meanwhile, is an online method to cope with motion segmentation and reconstruction simultaneously. They attempted to estimate the number of moving objects by splitting and merging feature tracks. They focused on practical considerations to build a reconstruction system, and their method can handle more realistic scenes.

Both [1] and [22] rely heavily on the feature tracker. Since Lucas and Kanade have introduced the KLT tracker based on optical flow [24], it has been used as a popular element for approaches to motion segmentation including methods for non-rigid object motions [25], [26]. In general, however, the feature tracker is prone to failure when the inter-frame motion is large. Given that the tracker assumes strong continuity between frames, one missing track can lead feature trajectories to drift away and make their return very difficult. Moreover, if the input data is given as an unordered set of images we cannot apply tracking-based methods to match the images.

Another issue of multiple object reconstruction is the relative scale between objects. If each object is treated separately, each 3D model may have a different scale factor because reconstruction is determined up to the unknown scale. This subtle problem is known as relative scale ambiguity problem, and has no exact solution; however a previous work [27] has introduced an estimation method based on generic motion constraints. In this method, the objects are re-arranged by their determined scale to create a whole 3D scene configuration.

Despite previous attempts to attain feature-level motion segmentation based on tracking method or theoretic factorization, the dynamic multi-object reconstruction problem has never been directly investigated. As a consequence, there is no practically available solution to this challenging problem. Thus, we emphasize that our work has made a substantial contribution in the form of a novel integrated system based on the object-centered approach.

3.3 Overview

In this section, we introduce the proposed reconstruction system, which uses an object recognition-based matching and segmentation. The main goal of a dynamic scene reconstruction system is to search for all major independent objects in multiple images and build 3D models of each object by gathering all visual information extracted from the images. The proposed approach stands on a number of existing techniques to achieve this goal. The proposed method consists of four major stages. The starting point of our algorithm is the automatic recognition of common objects among multiple images using a co-recognition technique, in which the recognized objects with the same identity across images are clustered together. Each cluster is a collection of the same object regions, roughly segmented in all images with matched feature correspondences. Following the recognition and the clustering steps, we applied structure from motion (SfM) to each object to calibrate the virtual camera. The SfM yields sparse 3D points on the object surface as well as the corresponding camera matrices through the bundle adjustment optimization. Then, the roughly segmented result of co-recognition and sparse 3D points projected on the images are

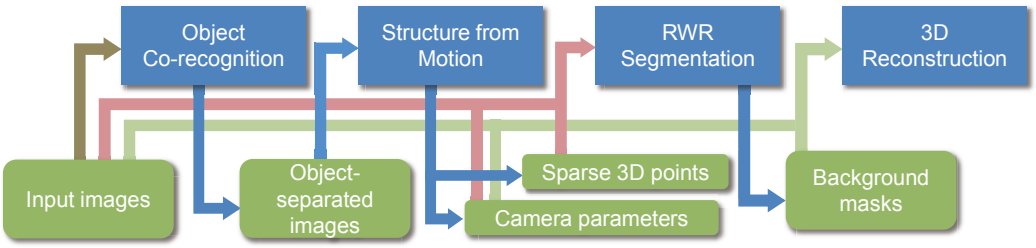


Figure 3.2: Overview of the proposed system

used as seeds for the RWR (Random Walk with Restart) segmentation algorithm. Finally, we used patch based multi-view reconstruction algorithm to build the 3D models of each object. An overview of our system is illustrated in Fig. 3.2.

3.4 Recognition

The first step of our algorithm is object recognition across images. Given that we have no object-level prior knowledge of the scene, we start by determining the number and locations of the objects in images. Given a set of images, as shown in Fig. 3.4, multiple objects appear simultaneously in a view, while their poses and arrangements vary in every view. To reconstruct multiple objects in the dynamic scene, each object should be segregated individually. By segmenting each object regions and finding the association among those in multiple images with feature correspondences, we can apply SfM algorithm on each separated object to construct its 3D model.

3.4.1 Co-Recognition

The object recognition problem for multiple object reconstruction is different from the exemplar-test object recognition framework. In our problem, there is no explicit distinction between exemplar and test images. Instead, we only have input images containing multiple dynamic objects. Therefore, the problem is to find and localize common objects in the images. The essence of our object recognition routine is based on co-recognition [28]. Co-recognition is an image matching method, which establishes correspondence among multiple common objects in image pairs without prior knowledge of the objects.

The building block of co-recognition is pair-wise image matching. Thus, we divided the matching problem of multiple images into sub-problems. Suppose we are given a set of N images, then we have a total of $N(N-1)/2$ image pair combinations. We solve each pair-wise sub-problem and get the final solution by integrating the results of those sub-problems.

Generative model:

We proceed by describing the generative model formulation. Given the image pair (I_i, I_j), the co-recognition problem can be modeled as a maximum a posteriori (MAP) estimation of the parameter θ on the basis of the image pair observation. According to the Bayesian formulation, the solution is found by maximizing posterior probability that is decomposed into likelihood and prior terms as follows.

$$\theta^* = \arg\max_\theta p(\theta \mid I_i, I_j) = \arg\max_\theta p(I_i, I_j \mid \theta) p(\theta). \tag{3.1}$$

We define θ as a set of K matching clusters μ between an image pair expressed

as:

$$\theta = \{\mu_1, \mu_2, ..., \mu_K\}. \tag{3.2}$$

Each matching cluster is a set of local patch matches between images given by:

$$\mu_k = \{\lambda_{k;1}, \lambda_{k;2}, ..., \lambda_{k;L_k}\}, \tag{3.3}$$

where L_k denotes the number of local region matches in μ_k. Thus, θ is a set of local correspondences grouped into the object-level matching.

The prior $p(\theta)$ represents the geometric properties that true common object matches should obey. It constrains the position of the matched local patches in the image pair. The relative position of the local patches in each cluster must have similar arrangement in both images. We penalized the position discrepancy between corresponding patches in the image pair. The geometric discrepancy error of cluster μ_k is the sum of local deformation cost of each local match, formulated as

$$\mathrm{E_g}(\mu_k) = \sum_{i=1}^{L_k} \mathrm{d_g}(\lambda_{k;i}). \tag{3.4}$$

The local deformation cost $\mathrm{d_g}(\lambda_{k;i})$ is the average geometric distances between the center position of neighboring patches of $\lambda_{k;i}$ and their matches in the other image. The neighbor relation is determined by the Delaunay triangulation of the patch centers, and the distance is measured in the normalized domain, in which the elliptical patch is transformed to the unit circle shape.

It also encodes the preference for larger clusters because reliable common objects are expected to have strong support from many local patches,

$$\mathrm{E_m}(\theta) = \sum_{k=1}^{K} \left(-L_k - |\Delta_k|\right), \tag{3.5}$$

where $|\Delta_k|$ denotes the number of Delaunay triangles of the cluster μ_k.

The likelihood $p(I_i, I_j|\theta)$ reflects the photometric similarity between matched patches in images. The normalized cross-correlation(NCC) values of the matched image patches are accumulated to measure the photometric error,

$$\mathrm{E_p}(\mu_k) = \sum_{i=1}^{L_k} \left(1 - \mathrm{NCC}(\lambda_{k;i})\right)^2 . \tag{3.6}$$

To sum up, we finally get the posterior probability from the prior and the likelihood with balancing parameter $\beta_\mathrm{p} = 3$ as

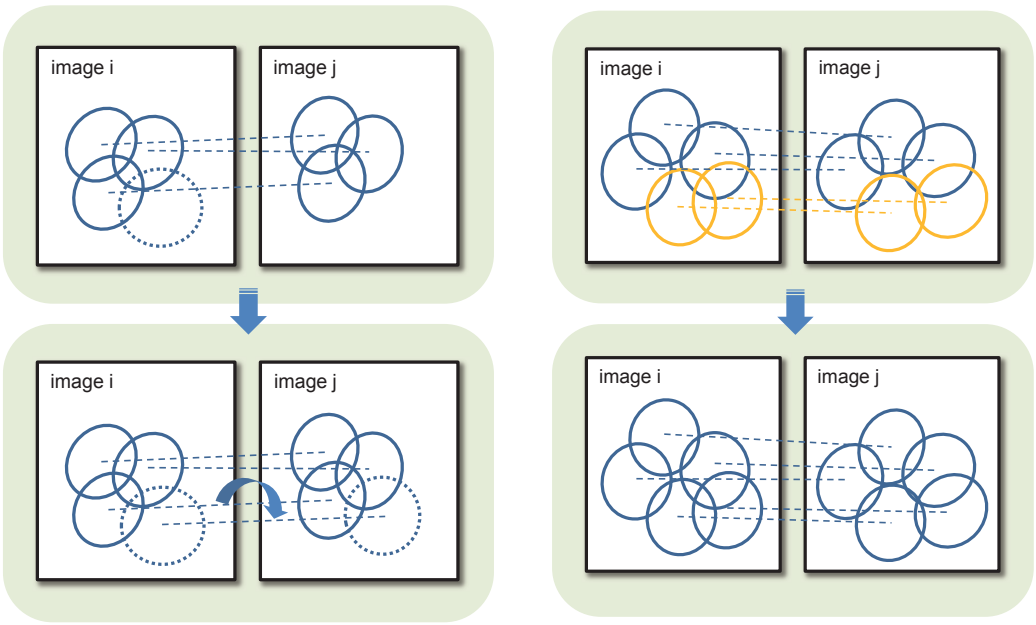
$$p\left(\theta|I_i, I_j\right) \propto \exp\left(-\sum_{k=1}^{K} \mathrm{E_g}(\mu_k) - \mathrm{E_m}(\theta) - \beta_\mathrm{p} \sum_{k=1}^{K} \mathrm{E_p}(\mu_k)\right). \tag{3.7}$$

Inference:

The first step of co-recognition is establishing initial feature-level matches between extracted local feature points. The feature detectors [29] [30] extract affine invariant regions from the images. Features with a distance of less than 0.45 in the SIFT descriptor space are considered to be matched. Usually, there are many false matches among initial matches.

After the initialization step, each feature match forms an initial cluster and grows to a larger cluster. Each initial cluster has its own expansion layer consisting of a set of overlapping circular grid, which covers the image. The overlapping circular grid on the image is the basic unit of growth. Then, we begin to run two iterative moves (expansion/merge) to grow the initial clusters.

In an expansion move, propagation and refinement operations are performed. The algorithm makes a proposal to propagate one of the current established matches to one of the unoccupied regions of the expansion layer. Then the new match is refined



(a) Expansion move

(b) Merge move

Figure 3.3: Expansion and merge moves. The concept of local patches and their correspondences between image pair (I_i, I_j) are expressed in elliptical regions and dashed lines connecting them. (a) In expansion move, current established local matches propagate an unoccupied region (dotted region) by transferring the transformation information from the nearby match. After local search to refine the propagated region, the new match is established and added to the cluster. (b) In merge move, two different clusters (depicted in dark blue and orange colors, respectively) merge into one cluster. This figure is best viewed in color.

by local search around the proposed region to find the best matching region. In a merge move, two large clusters are selected and merged into one. Also their expansion layers are combined. Figure 3.3 shows conceptual illustrations explaining the notion of expansion and merge moves. Expansion and merge proposals are accepted when the proposed state yields improved posterior $p(\theta|I_i, I_j)$. Expansion moves encourage merge moves to find congruous clusters by enlarging them. Likewise, merge moves help plausible expansion moves to have more expansion opportunities through gradual integration of compatible clusters. Utilizing cooperative expansion/merge moves, our greedy algorithm explores the solution space iteratively. Iterative growing is then performed until the convergence of posterior probability $p(\theta|I_i, I_j)$. After convergence, we eliminate unreliable clusters from θ^*. We measure the reliability of the cluster as the expanded area of the region, because larger clusters are more likely to originate from reliable seed matches.

Clearly, co-recognition has advantages over the feature tracking-based methods in terms of object identification. First, unlike the feature tracker, co-recognition-based approach can handle sudden object viewpoint changes between images efficiently. Given that inference starts from initial local feature matching, objects appearing at arbitrary position of images are recognized regardless of motion continuity. Although [1] and [22] can also reconstruct 3D model of multiple objects, they require smoothly captured video data, since they are based on the feature tracker. Besides, there are some prior works [31, 32, 33, 34] that perform reconstruction from unordered images; however, they require the scenes to be static. Thus, we argue that we solved more challenging and generalized problem to deal with unordered set of images containing

multiple dynamic objects. Second, due to expansion procedures of co-recognition, new matches are augmented from initial matches to cover object region. Therefore, the detected object regions are not restricted to the output of local feature detectors. Third, refinement presents flexibility to the expansion procedure and non-planar 3D objects are successfully recognized. It can be explained by the ability to overcome deformation coming from viewpoint variation.

3.4.2 Integration of the Sub-Results

As mentioned earlier, we divided the co-recognition problem on multiple images into sub-problems. The result of each sub-problem is a set of commonly appearing object regions matched in image pair. As depicted in Fig. 3.4, each object in one image can have several matching regions produced from pair-wise matches with different images in the dataset. Since each pair-wise matching is performed independently, there is no inter-connection of object identity between the results. Therefore, we combined the results of sub-problems into one integrated result. The integrated result has object-level correspondence network information.

The hierarchical agglomerative clustering [35] is used to unite pair-wise results. In the present work, we define the similarity measure of two object-level correspondences as the ratio of overlapping areas to the smaller region. Assuming that two matching clusters μ_p and μ_q, have a common sharing image, we let R_p and R_q be the region occupied by μ_p and μ_q on the common image, respectively. The distance between μ_p and μ_q is expressed as follows:

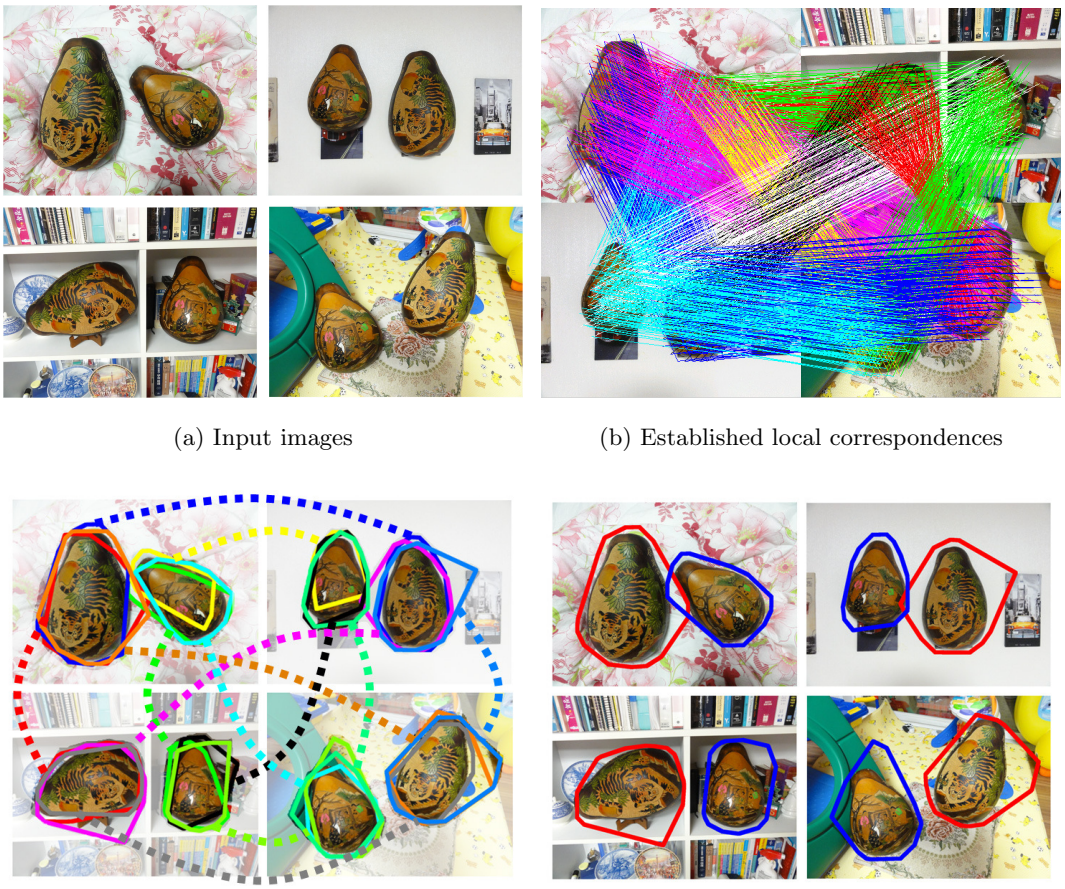
$$\mathrm{dist}(\mu_p, \mu_q) = \left(\frac{\min\left(\mathrm{Area}(R_p), \mathrm{Area}(R_q)\right)}{\mathrm{Area}(R_p \cap R_q)} \right). \tag{3.8}$$

The distance is set to infinity when there is no common image between the matching clusters. The object correspondences are joined by single-linkage hierarchical clustering until the distance is larger than 1.25. Object correspondences with distance closer than 1.25 are gradually agglomerated in the bottom-up manner. Therefore, the integrated result consists of detected object regions categorized into set of identical objects in the images.

As explained in Section 3.4.1, the basic unit of multi-layer growing is based on [28]. Although [28] detects identical objects within and across the images, we made a variation on it from a practical application aspect. Our algorithm applies different matching scheme according to each different type of data. The matches are allowed to be established in one of three ways: across the images, within the images, or across the temporal order. This modification increases efficiency and makes the algorithm applicable to various input data. The modification detail according to the input data type is described in Section 3.8.

3.5 Camera Calibration

The next stage in our system is camera calibration. After the recognition stage, all objects in the images were detected and clustered to the object correspondence network. Each set of object region segments across images satisfies the scene geometric consistency. They are equivalent to images of underlying object only taken by cameras in various positions and viewing angles, in which other distracting objects do not appear. Figure 3.4 (b) shows a typical example of object-centered segmented images. We perform camera calibration on each separated set of object regions depicted



(a) Input images

(b) Established local correspondences

(c) Pair-wise object-level matching

(d) Final integrated result of (c)

Figure 3.4: Pair-wise co-recognition and integrated result are illustrated on the *Gourd* dataset. (a) Input images have two objects in four different backgrounds. (b) The local correspondences are established between the instances of same object. (c) The pair-wise object-level matching and their resulting regions are superimposed on the images in different colors. (d) The identity of each object is distinguished by integrating the results of pair-wise object-level matching. Each color (red/blue) identities each object.

as blue and red outlines.

In this step, we compute the camera projection matrices of each individual object using the SfM technique. The point correspondences are optimized by the bundle adjustment optimization and the SfM yields both sparsely reconstructed set of 3D point coordinates $\mathbf{X}$ with camera matrices $\mathbf{P}$. Assuming we are given K objects in N images, we can have following set of camera matrices and points:

$$\{\mathbf{P}_{nk}; n = 1, ..., N, k = 1, ..., K\} \tag{3.9}$$

$$\{\mathbf{X}_k; k = 1, ..., K\}. \tag{3.10}$$

Note that not all objects have to be visible in all images; thus, if an object is missing in some images, the corresponding camera matrices and 3D points are not available.

Co-recognition produces object boundary segmentations up to the overlapping circular grid. Some extra parts outside the objects are included in the object regions due to the expanding nature of co-recognition. However, these small noises hardly affect the performance of SfM. The SfM multi-view constraint easily prunes these noises.

3.6 Object Boundary Refinement

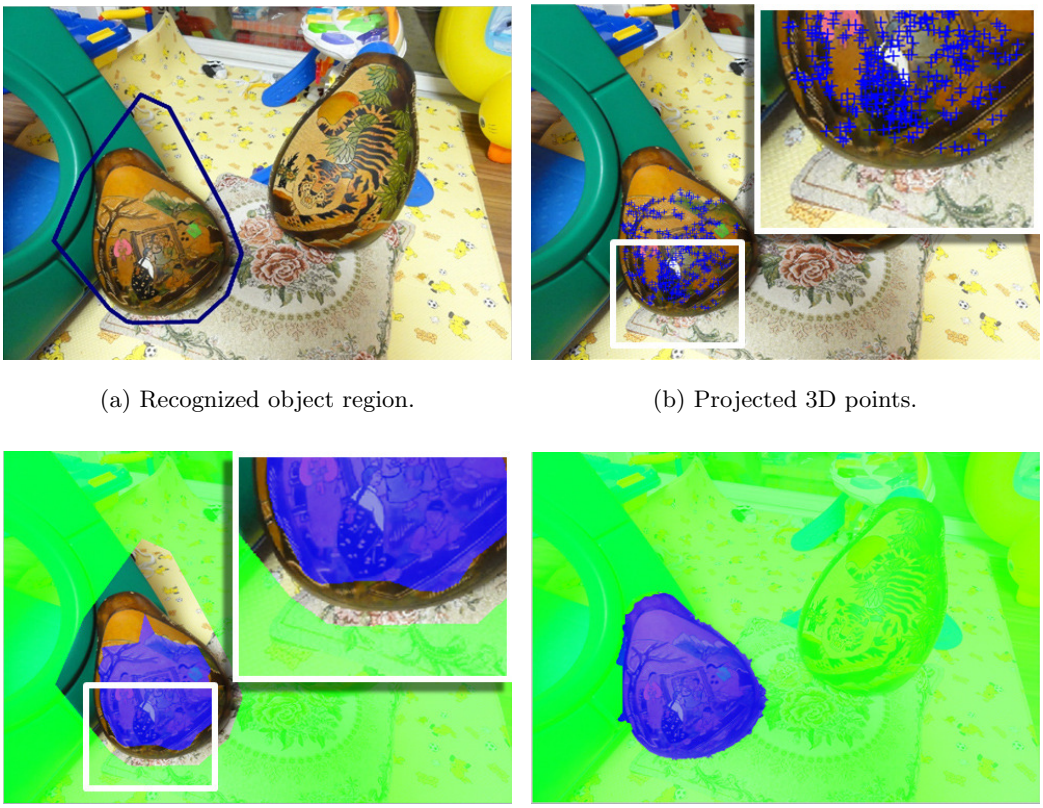
By virtue of object co-recognition, we perform object-centered camera calibration with segmented object images. Although the object boundary provided by co-recognition is useful in calibrating camera parameters, it is still rough and inappropriate for accurate 3D reconstruction of an object shape. It is clear that better object masks

enhance 3D reconstruction results by preventing unnecessary background parts from being processed. In this section, we apply the image segmentation method to obtain a detailed object segmentation boundary. We adopt the seeded segmentation method proposed by Kim et al. [36]. It is a generative image segmentation algorithm based on the Random Walks with Restart (RWR), and can efficiently solve the weak boundary problem and texture problem.

First, we construct a weighted graph in an image. The graph consists of pixel nodes and the edges connecting the neighborhood pixels. The edge weights encode image color similarity between connected nodes. Then, the random walkers traverse the graph with the probability proportional to the weights on the edges. We compute the steady-state probability for every pixel that a random walker starting at a seed point stays at the pixel. Finally, the most probable label is assigned to each pixel.

The RWR algorithm requires initial seed points for segmentation, and for this, the user provides scribbles as starting pixels of each label's random walker on the weighted graph. For binary labeling between object and background, seeds on the target object and background are required. Unlike interactive segmentation method[36], we aim to generate seed points automatically using RWR as a non-interactive segmentation method.

The key idea behind providing reliable seed points is utilizing an object's geometric information. Although discontinuity of visual pattern is typically observed at the object boundary, sometimes it is ambiguous to decide whether a pixel is on the object or not by the photometric observation only. As explained in Section 3.5, the sparse 3D points as well as the camera matrices are extracted by SfM under the



(a) Recognized object region.

(b) Projected 3D points.

(c) Object and background seeds.

(d) Refined object boundary.

Figure 3.5: Object boundary refinement by RWR segmentation. (a) The object boundary provided by co-recognition is marked in blue line. (b) Blue crosses denote 3D points projected on the image. White box is scaled for display. (c) Blue mask represents *object* seeds produced by alpha shape, and green mask denotes *background* seeds given by object recognition. (d) Refined object region after proposed non-interactive RWR segmentation.

consideration of 3D geometry. This implies that projected locations of sparse 3D points on an image plane are most likely to lie on the object's surface. The sparse 3D points $\mathbf{X}_k$ of object k are projected on the image n by the projection matrix $\mathbf{P}_{nk}$ as follows:

$$\{\mathbf{x}_{nk} = \mathbf{P}_{nk}\mathbf{X}_k; n = 1, ..., N, k = 1, ..., K\}. \tag{3.11}$$

In addition to the projected points $\mathbf{x}_{nk}$, we apply 2D alpha shapes [37] to them to obtain more stable seeds for segmentation. The alpha shape is a polygon derived from the point set with parameter α controlling the desired level of detail. The alpha shape generated from $\mathbf{x}_{nk}$ fills the empty space between the projected points. The seed is now given as a polygon area instead of each projected point. We empirically determined the optimal α value to adapt to the scale and texture density of the objects by following procedure. We select the 6 nearest neighbors of every points and calculate the average distance from the selected points. The α value is set to twice the average distance.

For the background seeds, we simply mark all the points outside the object boundary obtained from co-recognition. Figure 3.5 shows the segmentation process with the automatically generated seeds. The object boundaries are determined to pixel-level precision through the RWR segmentation stage.

3.7 3D Reconstruction

The final stage of our system aims to reconstruct a 3D model for each object. Given object segmentation masks and dense correspondences with camera projection ma-

trices, the condition for running 3D reconstruction algorithm is satisfied. One can use any reconstruction method to obtain 3D models of objects and background. In the paper, we adopt the publicly available multi-view stereo software PMVS [38], considered as the state-of-the-art algorithm.

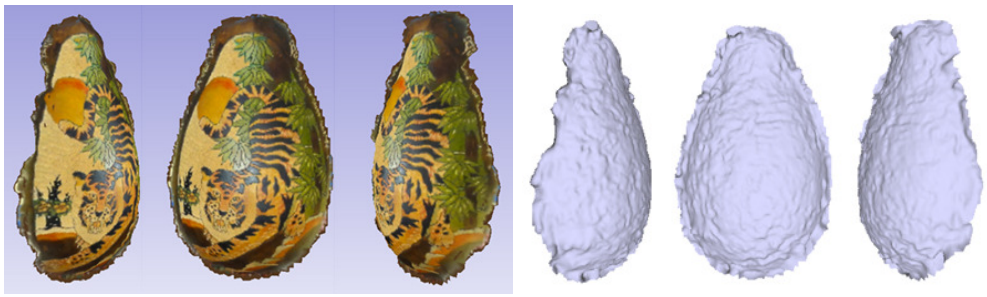
3.8 Experiments and Results

In this section, we demonstrate the experimental results of our approach for multiple object reconstruction in dynamic scenes. We demonstrated the performance of our algorithm on several test image sets containing objects that exhibited varying geometric configuration across frames, on both different and same backgrounds. For the experiments on video data, video sequences captured from movies and video clips downloaded from the Internet are used, as well as the video taken in the Lab. Comparisons are drawn with some prior works that have similar goals with our approach. We also performed quantitative evaluations of pixel-wise segmentation accuracy and 3D reconstruction correctness. To show the process and result of our approach more effectively, we uploaded the supplementary video material on our web site: `http://cv.snu.ac.kr/research/~MORDS/video_MORDS.wmv`

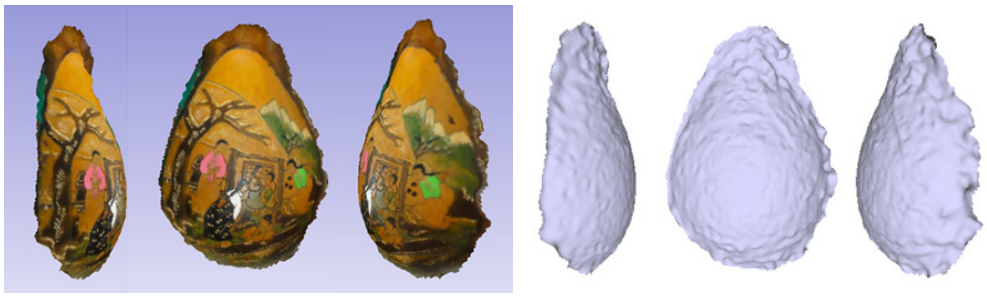
3.8.1 Qualitative Results

Dynamic scenes with different backgrounds:

We first performed experiments on the sets of images, which capture multiple target objects on different backgrounds in each shot. We took the *Gourd* and *Tea* dataset which are comprised of 4 images as shown in Fig. 3.4 and Fig. 3.7(a), respectively.



(a) *Gourd* object 1



(b) *Gourd* object 2

Figure 3.6: Reconstruction results of two objects of *Gourd* dataset.

(a) Input images contain three objects in different configuration and background.

(b) Reconstruction results of three objects.

Figure 3.7: *Tea* dataset and the reconstruction results.

Our goal is to reconstruct the common objects which appear in all images. The scene continuity between consecutive frames is a crucial assumption that the tracking-based methods rely on. However, in this setting of experiment, every image has its unique configuration of scenes. Any permutation of input images will yield abrupt change of object poses and positions. The target objects are shown in various poses and positions in the scenes. Given that they have no consistent background, our algorithm utilized visual information from the foreground objects only. Figures 3.6 and 3.7(b) show the reconstruction results of *Gourd* and *Tea*, respectively. Considering that the target objects occupy a small part of the images and only 4 images are used to reconstruct the 3D models, the effectiveness of our system is quite convincing.

Dynamic scenes with constant background:

The second experiment is designed for reconstructing the foreground objects and background parts. For this purpose, we captured *Houses* dataset. Figure 3.8(a) shows some sample images from the *Houses* dataset. *Houses* is a sequence of 25 images. Two objects have independent abrupt motions with a consistent background, while the camera moves left and right. Since the images contain consistent background, our algorithm separates the foreground and background by detecting them as distinct objects. The 3D model of the background part is also reconstructed as well as the foreground objects. Figure 3.8(b) and 3.8(c) show the reconstructed background and full 3D shape of each object, respectively. The explicit segregation of the object and background has enabled reconstruction of occluded background. Due to the occluding objects, some part of the background is not seen from the camera's view, however, images taken from other viewpoints compensate for the missing part. Note

that this is different from the crowded scene reconstruction presented previously [38] [39], which treat occluding objects as obstacles that have to be filtered out.

Identical objects in one image:

Interestingly, the proposed method is applicable to the reconstruction from a single image if the single image contains multiple shots of identical objects. Fig. 3.9 shows our example of *Milk*. The repeated visual pattern induced by the multiple instances of identical object is frequently observable in the real world. Each object region in the image is equivalent to each shot of same object taken from different viewpoint. The relative camera position varies as the objects have different poses seen by the single camera.

Here, we decompose each instances of the object in the recognition stage, and they are treated as multiple shots of same object. In such a case, the image matching is done only within the single image itself. To perform the single image reconstruction, we carry out a little modification to the initialization step of co-recognition. We allow the local features to find initial correspondences from the feature points extracted from the same image. The self-matched regions grow to all of the identical object regions. As shown in the reconstruction result in Fig. 3.9, our method provides good reconstruction result from a single image.

Dynamic scene in video clips:

In the fourth experiment, we performed experiments on the *Racing* and *Dolls* video. We captured the *Racing* video from YouTube, and the *Dolls* was taken in our Lab.

In the *Racing* video clip, the camera is fixed and the viewpoint does not change. Instead, two cars appear and disappear in the sequence as they move across the

circuit. This video scene contains consistent background, but it does not have relative camera motion. Although our system runs without image ordering, we exploit the ordering information of video frames to increase efficiency and overcome the high computational complexity of matching all possible pairs from the combinations. The frame at time t is matched with the frame at time $t+3$ and $t+6$ sequentially. A total of 36 frames were used in the experiment. The reconstruction result is shown in Fig. 3.10. Despite the blurry low texture of the body and window glass, the two cars are separated and modeled to 3D shapes successfully.

In the *Dolls* video clip, two dolls revolve around each other in the background. This scene is captured by a moving camera. The two objects occlude each other when one is located between the other and the background, then the occluded object re-appears as the revolution continues. We observe partial/full occlusions of the target objects in the video. This video raises several challenging issues such as scale change, treatment of consistent but occluded background, mutual occlusion of objects, and re-identification of disappeared object. We apply the same strategy used in *Racing* video to match the *Dolls* sequences. A total of 160 frames are used in the experiment. Figure 3.11 shows the reconstruction results of two doll objects and the background. In this experiment, we also show that the hidden occluded parts of the target object can be estimated by utilizing the built 3D model. As shown in Fig. 3.11(d), if the object undergoes partial occlusion, and the remaining observable part of the object provides enough visual information to calculate the object pose in the frame, the occluded part can be estimated. Since the information of the object shape is aggregated from images taken in various viewpoints, the knowledge of

the partial visual information can lead us to the whole object structure throughout their reconstruction procedure. The results reveal that proposed method successfully overcomes the aforementioned challenges, which are known to be difficult issues of conventional methods.

Comparison with [1]:

We performed an experiment on the same video in [1]. A scene where a van moves on the road as the camera pans right was extracted from the movie *Groundhog day*. A total of 30 frames were used in our experiment. Rothganger et al. [1] modeled the object as a set of affine covariant surface patches extracted from the feature detectors. As their algorithm runs only on the given patches, the reconstruction result is limited to the sparsely and unevenly distributed interest regions induced by the detectors. However, our algorithm explores new regions, which were not included in the initial output of detectors, through the expansion process. The results of our method and [1] are displayed in Fig. 3.12 for comparison.

Comparison with [2]:

To contrast robustness with abrupt motion, we compared the object detection performance of our method with the work of Wills et al. [2]. Their method was intended to overcome the large inter-frame motion disparity, which was the biggest problem of tracking-based motion segmentation algorithms. We selected an image pair from *Tea* and applied both methods on the same image pair[1]. The object boundary determined by our method is displayed in Fig. 3.7(b) and Fig. 3.7(a) reports the failure of [2]. We got similar results from any combination of image pair. Although the

[1]We used the code provided by the authors (`http://joshwills.com/projects/www_code.html`).

method in [2] has been developed to handle large abrupt motions, their approach is weak in terms of scale change and rotational transformation. They have not explicitly modeled local affine transformation between the true inlier matches. Moreover, their assumption of planar motion is not appropriate for 3D objects. When a 3D object has out-of-plane rotation, each of its local region undergoes different movement. Our algorithm adapts to the deformation of surfaces as well as arbitrary positioning of objects.

3.8.2 Quantitative Results

3.8.2.1 Evaluation of Segmentation Accuracy

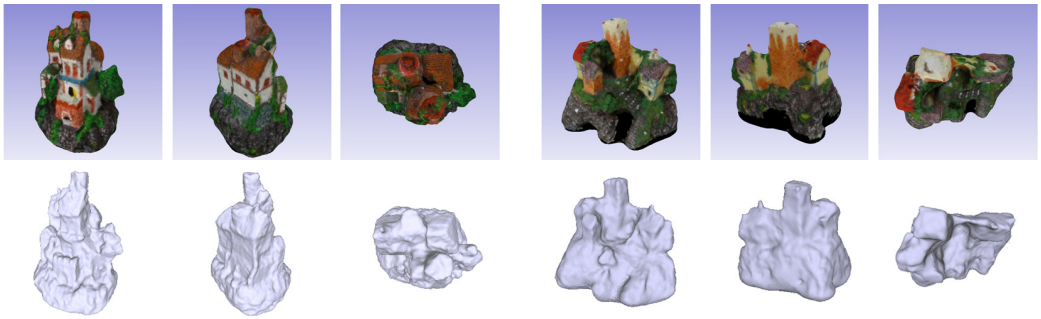
As a matter of quantitative evaluation, we measured the segmentation accuracy of target objects after each step of co-segmentation, RWR, and 3D reconstruction. For co-recognition and RWR, detected regions of the objects were compared with ground truth. For 3D reconstruction step, we then re-projected 3D models on 2D image planes to obtain the segment.

To measure segmentation performance, we manually labeled the target object region's ground truth pixels. We measured the segmentation accuracy by three criteria. The hit ratio was calculated as the ratio of truly detected pixels to the ground truth pixels, $HitRatio = |Result \cap GT|/|GT|$. The background ratio refers to the ratio of false positive pixels to result pixels, $BkgRatio = |Result - GT|/|Result|$. The overlap ratio measures the degree of overall correctness of segmentation, as the ratio between intersection and union of the result and ground truth, $OverlapRatio = |Result \cap GT|/|Result \cup GT|$. The higher hit, overlap ratio and lower background

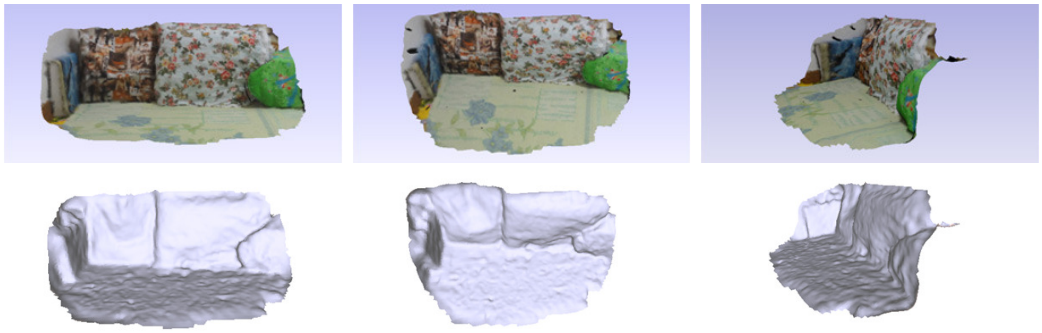
ratio means we have obtained better segmentation results.

We took the measurements of the *Gourd, Tea,* and *Milk* dataset, which consist of image shots. Their results are shown in Tables 3.1, 3.2, and 3.3, respectively. Also they are visualized as graphs in Figure 3.14. As shown in the Tables, the RWR segmentation step has significantly increased the segmentation accuracy. The tendency has shown that co-recognition detects relatively larger regions than the target object. This result explains that co-recognition detects each object region as a cluster of overlapping circular grid, which has expanding properties, while RWR finds pixel-level object boundary segmentation. The expanding nature of co-recognition yields high hit ratio, but also increases background ratio. The RWR mostly refined the object boundary by filtering out the background part. The RWR segmentation decreased the background ratio with little degradation of the hit ratio. Interestingly, the 3D reconstruction step sometimes shows slight decrease in the measured segmentation accuracy in terms of intersection ratio: 0.94 to 0.93 for *Gourd* and, 0.97 to 0.95 for *Milk*. The reason for this is that 3D reconstruction uses relatively conservative criteria to model objects. For the reliability of 3D model, part of an object is reconstructed when it can be seen at least three different views. The small degradation of reprojected 3D reconstructed model is easily explained by the fact that ground truth segmentation is made solely along the object boundary observed in each view.

(a) Sample images from 25 input images of *Houses*.



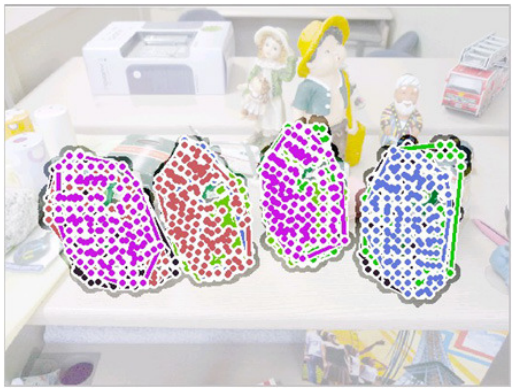
(b) Reconstructed 3D model of each object.



(c) Reconstructed background

Figure 3.8: *Houses* dataset and reconstruction result of two objects and background.

(a) Input image of *Milk*.



(b) Detected object instances as sets of local regions.

(c) Reconstructed 3D model.

Figure 3.9: Reconstruction of identical object from a single image.

(a) Sample frames from *Racing* video clip.

(b) Reconstruction results of two racing cars.

Figure 3.10: *Race* video clip and the reconstruction results of two cars. Note that the consistent background part is not reconstructed since the camera is fixed.

(a) Sample frames from *Dolls*.

(b) Reconstruction results of two dolls.

(c) Reconstruction result of background part.

(d) One object is partially occluded by the other object. We can estimate the occluded part of the object by utilizing the built model. Left: input frame. Right: 3D model of partially occluded object superimposed on the image.

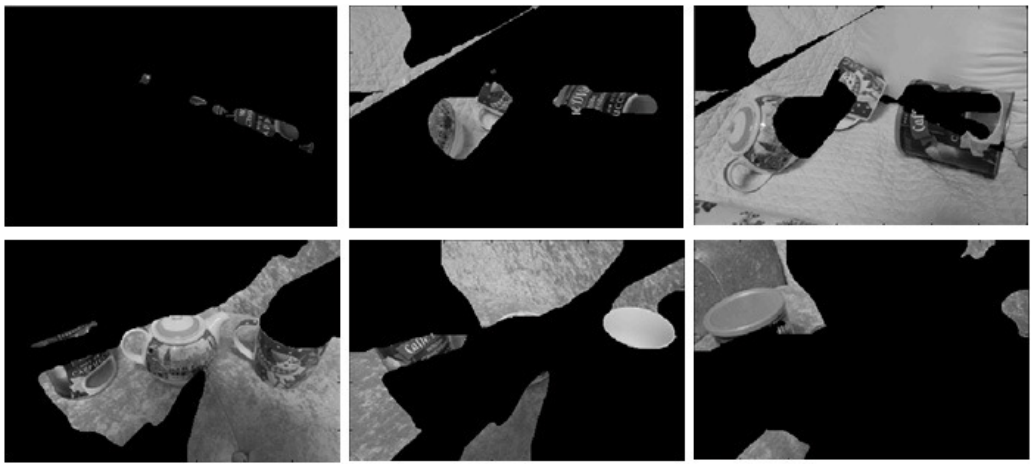
Figure 3.11: *Dolls* video clip and reconstruction results.

(a) A frame from *Groundhog day*.

(b) Reconstruction result of [1].

(c) Our reconstruction result.

Figure 3.12: Experimental results on a scene of the movie *Groundhog day*. Our approach shows better result then [1].



(a) Result of [2], where each column corresponds to each object.

(b) Our result, where each color represents identity of each object.

Figure 3.13: Comparison of object detection performance with [2] reveals the superiority of our method in abrupt motion of 3D objects. Detected object boundary on an image pair of *Tea* is displayed.

Table 3.1: Segmentation performance (*Gourd*)

	Object 1			Object 2			Total		
	Hit	Bkg	Overlap	Hit	Bkg	Overlap	Hit	Bkg	Overlap
After Co-recognition	.998	.309	.690	.994	.272	.727	.996	.295	.704
After RWR segmentation	.966	.017	.950	.965	.025	.941	.965	.020	.946
3D model reprojection	.961	.016	.946	.938	.034	.908	.952	.023	.931

Table 3.2: Segmentation performance (*Tea*)

	Object 1			Object 2			Object 3			Total		
	Hit	Bkg	Overlap	Hit	Bkg	Overlap	Hit	Bkg	Overlap	Hit	Bkg	Overlap
After Co-recognition	1.000	.725	.275	.998	.553	.446	.999	.460	.540	.999	.591	.409
After RWR segmentation	.990	.105	.887	.962	.123	.848	.965	.100	.871	.971	.107	.870
3D model reprojection	.979	.018	.962	.899	.023	.880	.900	.017	.886	.922	.019	.906

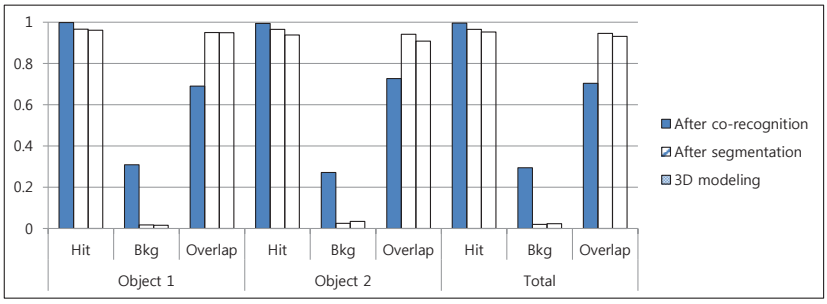
Table 3.3: Segmentation performance (*Milk*)

	Object 1		
	Hit	Bkg	Overlap
After Co-recognition	.992	.318	.679
After RWR segmentation	.997	.070	.972
3D model reprojection	.972	.016	.957

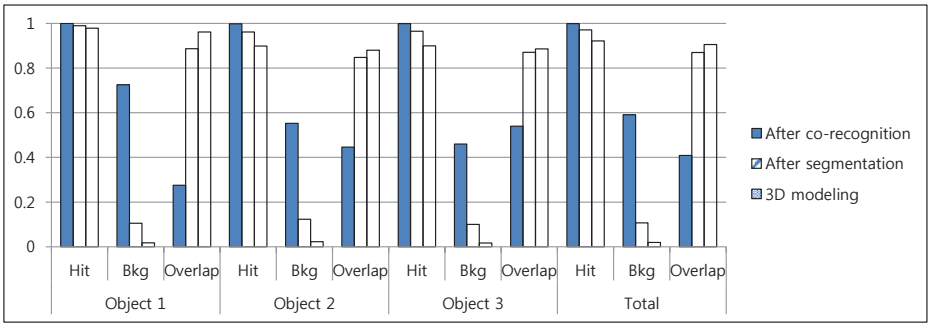
3.8.2.2 Evaluation of Reconstruction Accuracy

It is difficult to evaluate the reconstruction accuracy without the absolute, dense ground truth of the 3D object surface. However, some knowledge on the target scene's geometric relationships can be used to measure the accuracy of the built 3D model indirectly [22].

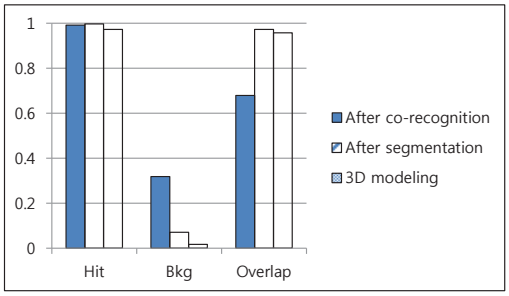
In our experiment, we used the measured angle between two perpendicular planes of an object as the reconstruction accuracy measure. As shown in Fig. 3.15, we took two different sides of the *Milk* object and fitted a plane on each of its surfaces in a total least square sense. The principle component analysis (PCA) was used to fit the plane models to the 3D coordinates of the points on each reconstructed planes. Assuming that the points constitute a plane, the coefficients of the 3rd principle component define the normal vector of the plane by the underlying theory of the PCA. The deviation of the angle between the normal vectors of the two reconstructed planes shown in Fig. 3.15 from 90° was measured as 4.92°. This shows fair accuracy in spite of each object instance's small occupancy in the single image and limited viewpoints.



(a) *Gourd*



(b) *Tea*



(c) *Milk*

Figure 3.14: Segmentation performance of *Gourd*, *Tea*, and *Milk* data.

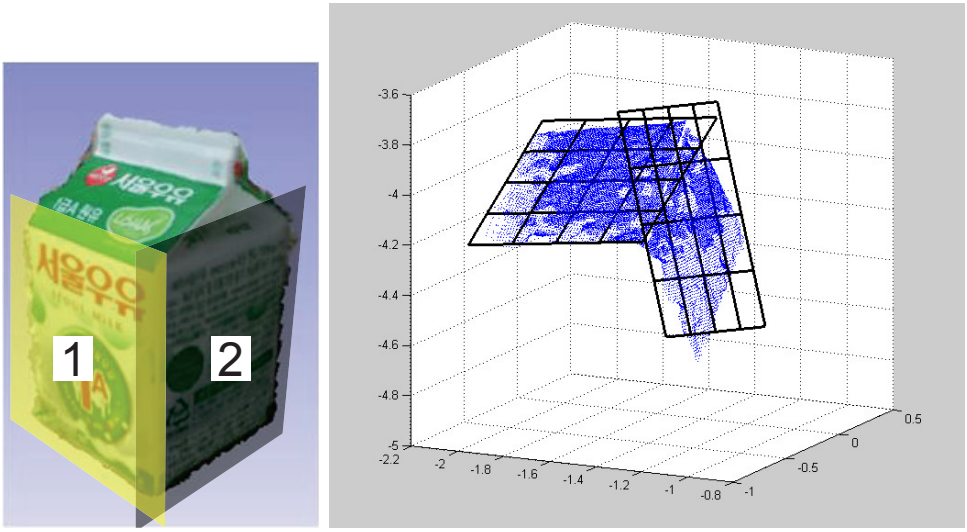


Figure 3.15: Perpendicular planes in the object are used to measure the accuracy of the 3D reconstruction. Two sides of the *Milk* object are fitted on two planes and their relative angle is calculated using normal vectors.

3.8.3 Analysis

Occlusion:

A feature of the proposed algorithm is robustness against object occlusion. Conventional approaches [1] [22] are based on the feature tracking method where the algorithm requires explicit treatment of occluded objects. Given that the tracked features show discontinuity at the occlusion boundary, the disconnected features have to be saved and restored to deal with the occlusion. However, our system recognized the object identities instead of following them. Thus, occlusion handling does not require additional explicit process. Furthermore, our recognize-integrate scheme enables us to deal with scenes containing full occlusion or missing object. Disappeared objects can maintain the same identities only if they keep similar appearance.

The *Dolls* video was intended to capture the severe occlusion situation. When an object passes behind the other object, the occluded object disappears completely from the camera's view. However, if an object is partially occluded as much as the calibration is available using the observable part, we can estimate the hidden part of the object by projecting the built object model. Figure 3.11(d) shows an examples of occlusion.

Computational complexity:

We performed the experiments on a computer with 3.3GHz processor. The co-recognition and RWR are implemented in Matlab, while SfM and dense 3D reconstruction are written in C++.

In our framework, the majority of computational time is devoted to the co-recognition and the dense 3D reconstruction steps. For instance, the co-recognition step took a total of 1,380 seconds to match the *Gourd* images. The *Gourd* data consists of 4 images ($N = 4$), yielding 6 pairs of image matching sub-problems. On average, the running time of each pair-wise sub-problem was 230 seconds with standard deviation of 23.1 seconds. The camera calibration step took 23 and 22 seconds for each object, totally 45 seconds to run the SfM algorithm on both recognized objects. The object boundary refinement step was performed within nearly 1 second for each image, consuming the smallest computation time in the whole pipeline. The dense 3D reconstruction step spent 143 seconds to build the 3D model of the first object, 103 seconds to the second object.

The computation time of the each co-recognition sub-problem varies according to

the foreground objects' occupied spatial regions in images. The algorithm converges relatively quickly in the case of small objects, since only small number of iterations are needed to find the object regions. On the other hand, larger objects require more time to be recovered by the iterative region growing process. For the *Race* dataset, more than 500 seconds were needed in each co-recognition sub-problem, since the region growing expands to the whole image region although the background part will not be used in later steps. The parameters that control the reconstruction density mainly determine the execution time of the dense 3D reconstruction.

Limitations:

Although our system presents a robust framework against challenges such as occlusion (*House, Racing*, and *Dolls*) or affine, perspective changes (*Tea*), it also has limitations. Empirically, the structure-from-motion step is the weakest part of the flow. Objects with planar or shallow-depth structure can raise degeneracy to SfM. In such cases, despite the success of object recognition step, the objects cannot be reconstructed. Next, the object recognition step requires sufficient texture on the surface of target objects since each local patches are matched by textures (i.e. SIFT descriptors, NCC measure). For instance, the teapot object of *Tea* images contains low-textured handle and lid, but only the body of the teapot is recognized. The texturedness also affects the quality of dense reconstruction. The reconstruction results in Fig. 3.10 and Fig. 3.12 show that blurry low-textured body and window glass are reconstructed ruggedly.

3.9 Summary

In this chapter, we have presented a reconstruction framework for multiple objects in dynamic scenes. We designed a system based on the object recognition approach, which solves the problem of estimating object number and feature matching issues at once. Our object-centered approach grounds on the fact that many of vision problems can be interpreted as correspondence problems. Thus, unlike the conventional flow-based approaches, the proposed method utilized the correspondence information acquired from the unsupervised co-recognition method. We presented our work as an integrated framework, which includes unsupervised object recognition, segmentation, and 3D model reconstruction from continuous or discontinuous dynamic scenes. Experimental results on various data have demonstrated the effectiveness of our approach, especially in the presence of abrupt motion of objects.

For our future work, we shall attempt to reconstruct a variety of scenes containing non-rigid objects with more complex motions. We hope this will eventually lead us to a practical technique, such as 3D conversion of old classic movies.

Chapter 4

Super Resolution for 3D Face Reconstruction

4.1 Introduction

The conventional definition of the *super resolution* is a technique to enhance the resolution of an image. This problem is a representative inverse problem in computer vision, which aims to recover the true information of the physical object from the insufficient observation. Due to its practical applicability and academic interest, various approaches have been introduced. However, the main subject of conventional super resolution technique are 2 dimensional intensity images, as they are abundant from many years ago and easy to represent in 2D matrix data.

In this chapter, a new method for super resolution of 3D data is proposed. The 3D data super resolution, as discussed in the paper of Aodha et al. [40], is different from the intensity image super resolution. Compared to the intensity image, the depth

image is less affected by the lighting condition or the texture of the surface, but the noise characteristic is very different from the intensity image. Thus, a different scheme was proposed for depth image super resolution [40]. However, in a strict sense, the depth image is considered as a 2.5D data in terms of shape representation. The depth image is a 2D matrix data, of which each pixel contains distance between the camera and the corresponding surface of object. Since the depth image contains the distance data acquired from the viewpoint of specific camera position, it is regarded as a camera-centered data. The point cloud data, unlike the depth image, contains viewpoint invariant geometrical structure information of the target object. Thus, in this chapter it is regarded as an object-centered data.

The main goal in this chapter is applying a super resolution to the 3D point cloud data. More specifically, the super resolution is applied to the low resolution single 3D model point cloud data, without additional depth image or intensity image observation. This chapter also concentrates on the human face as the target, which is the prominent interest of human visual perception system.

Compared to the image data, treating the point cloud data requires special care and additional efforts. For example, while image pixels can be one-to-one matched when they are overlapped on each other, the points in the cloud cannot find exact and trivial matching points since they are distributed freely in the 3D space. The difference between image and point cloud is visually exposed clearly when the are displayed on the digital computer screen. The camera-centered image data has no information about the physical size of the target. If the screen displays the image in actual pixel size, the low resolution image generally occupies smaller region unless

it is intensionally stretched by the interpolation method. However, in the object-centered point cloud data, the information of the geometric size of the target object is represented as the coordinates of the points, regardless of the cardinality of the point cloud. Thus, when a 3D point cloud model is down sampled to a lower resolution, the space occupied by the points does not shrink, instead, the *density* of points decreases.

To our knowledge, this is the first attempt to synthesize a super-resolved 3D model from a single low-density 3D point cloud data without additional depth observations or intensity information.

4.2 Related Work

There has been proposed various approaches and methods to intensity and depth image super resolution problem. In this section, the depth image super resolution methods are reviewed.

While many intensity image super resolution methods are proposed, relatively few researches have been done for the depth image super resolution as a counterpart of the intensity image super resolution. Because of the preceding development of intensity image super resolution, some depth image super resolution method are inspired by the intensity image method.

First of all, Schuon et al. fused multiple low-resolution depth images to generate a high-resolution depth image, which is analogous to the multi-frame approach for intensity image [41]. Also, Rajagopalan et al. [42] proposed an MRF formulation, on which multiple low-resolution depth images are combined to build a high-resolution

result.

Secondly, approaches utilizing the superior resolution of intensity images are proposed. Yang et al. [43] proposed a hybrid approach to up-sample a low-resolution depth image with the help of accurately registered high-resolution intensity image. Diebel and Thrun [44] applied multi-resolution MRF to merge range and intensity images in different resolution. Kiechle et al. [45] proposed a bimodal co-sparse model for inferring high resolution depth image from the low resolution depth image registered with high resolution intensity image. In the work of Ferstl et al. [46], the anisotropic diffusion tensor is calculated from the high resolution intensity image and applied to the low resolution depth image for high speed up-sampling. The intuition behind the depth-intensity image hybrid approach is that the discontinuity of the depth co-occur with the edge boundary in the intensity image. To apply this assumption to the real problem, a careful registration between depth and intensity images is needed.

Recently, Aodha et al. [40] suggested a single-frame depth image super resolution method. They used a synthetic images of 3D generic scenes as a training exemplar database, showing that severe noise in depth images can be reduced by applying artificial shape prior. Although the approach of Aodha et al. is similar to the exemplar based single-frame intensity image super resolution [17] in terms of MRF construction, they applied different procedure to deal with the depth images, which have different characteristics.

4.3 Overview

In this section, we introduce the proposed algorithm and the role of each component consisting overall procedure is briefly described. Given the low resolution 3D face input model, the basic strategy of super resolution is to 1) decompose the face into local patches, 2) find high resolution counterpart for each patch, and 3) re-construct the local patches into the single super-resolved output.

First, the proposed algorithm starts with building an exemplar database. The database contains information which is needed to up-sample the low resolution 3D data to the high resolution one. The database stores the information as a set of pairwise matches of descriptor and high resolution patches, which are learned from the training data. The learning procedure consist of two main steps. Extracting high resolution local 3D patches, and describing the down-sampled patches with the surface descriptor. Since the descriptors are used as the query key to the database, the exemplar database does not have to store low resolution patches. The learning process is conducted off-line.

At the run-time, the low resolution patches from the input data inquire of the database about the most similar patches of them in the database. Since the patches have different cardinality, the patches are described by a surface descriptor. The exemplar database outputs several high resolution patches for each low resolution input patch by comparing their descriptors. Through the optimization process, the retrieved high resolution patches are selected, and they are reconstructed as the final super-resolved result. Figure 4.1 shows the overall procedure of proposed framework. The orange arrows in figure 4.1 depict the off-line process, and the blue arrows mean

the processing flow in the run-time.

The proposed procedure allows for non-parametric super resolution, since no parametric knowledge about the relation of low-high resolution 3D data is required.

4.4 Proposed Model

In this section, the proposed model for the super resolution of the 3D point cloud face data is discussed. Given the low resolution 3D point cloud face model X, our goal is to produce a plausible high resolution 3D model Y from X. To represent the data, the Markovian property is assumed. The 3D model data is divided into local patches, and each of the patches is assigned to a node in the Markov Random Field [47], and they are written as:

$$X = \{x_1, x_2, ..., x_N\}, \tag{4.1}$$

$$Y = \{y_1, y_2, ..., y_N\}, \tag{4.2}$$

where N denotes the total number of nodes.

Two neighboring nodes in the Markov Random Field (MRF) are statistically dependent, and they are connected to each other with the pairwise compatibility function ψ. Each of the observable node x (low resolution patch) is connected to its corresponding latent variable y (high resolution patch) with compatibility function ϕ.

Thus, for the Markov Random Field, the joint probability of the observation X and latent variable Y is proportional to the product of all compatibility functions

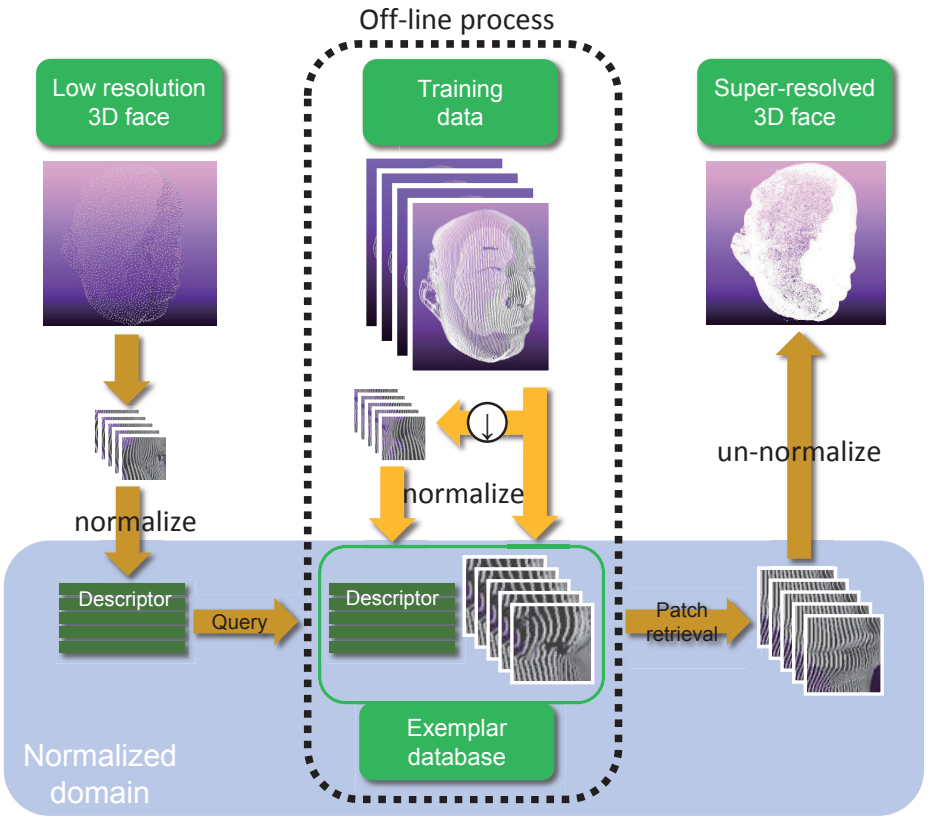


Figure 4.1: The overview of the proposed algorithm. In the off-line process, the exemplar database is constructed from the training data. Many local patches are extracted from the training data, and the matched descriptor of the down-sampled version of patches are stored in the database. The patches from the low resolution input find candidate high resolution exemplars by comparing the descriptors, and they are reconstructed to build a super resolution output.

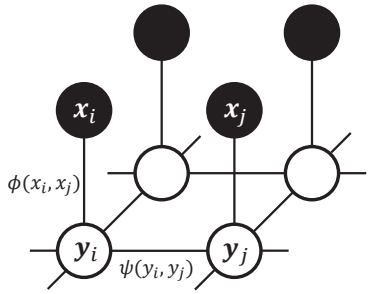


Figure 4.2: Graphical model representation of the Markov Random Field (MRF) for the proposed method. Each node in the graph means a local patch of the 3D face data. The observed variable x, which denotes a low resolution input patch, is colored in black. The latent (hidden) variable y, which denotes a high resolution patch, is colored in white. Connected line indicates statistical dependency.

of the network, as expressed in the following formula:

$$P(x_1, x_2, ..., x_N, y_1, y_2, ..., y_N) = \frac{1}{Z} \prod_{(i,j) \in \mathcal{N}} \psi(y_i, y_j) \prod_k \phi(x_k, y_k). \tag{4.3}$$

Figure 4.2 illustrates the graphical model representation of the Markov network used in the proposed method. The black circle node x means the local patch in observed data, and the white circle node y indicates the latent variable which should be estimated. The statistical dependency is denoted by a connecting line and compatibility function between nodes.

Here, we aim to estimate the uncovered variable Y as a solution of super resolution of 3D data. Given the low resolution X, the super resolution problem is modeled as a Maximum a Posteriori (MAP) estimation of the variable Y. The Bayesian formulation decomposes the posterior probability into prior and likelihood term as:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \propto P(X|Y)P(Y). \tag{4.4}$$

Thus, estimating the MAP solution is formulated as finding Y^* that maximizes the value of the following formula:

$$Y^* = \arg\max_Y P(Y \mid X) = \arg\max_Y P(X \mid Y)P(Y). \tag{4.5}$$

Alongside the aforementioned equation (4.3), the likelihood $P(X \mid Y)$ corresponds to the unary potential ϕ, and the prior $P(Y)$ corresponds to pairwise potential compatibility function ψ.

For the ease of computation, we follow the mathematical convention that takes the negative logarithm to the probability function. Based on the fact that each potential function is independently distributed, the logarithm turns the product of them into the summation, yielding the energy function. The energy function associates low energy to correct values and high energy to incorrect values.

$$\begin{aligned} E(Y|X) &= \sum_k -\ln \phi(x_k, y_k) + \lambda \sum_{(i,j)\in\mathcal{N}} -\ln \psi(y_i, y_j) \\ &= \sum_k E_d(x_k, y_k) + \lambda \sum_{(i,j)\in\mathcal{N}} E_s(y_i, y_j). \end{aligned} \tag{4.6}$$

4.4.1 Local Patch

The local patch is the basic unit of producing super resolution. Extracting the local patches in the point cloud is more difficult than extracting patches from the matrix image. In the conventional image, of which the pixels are arranged in 2D regular

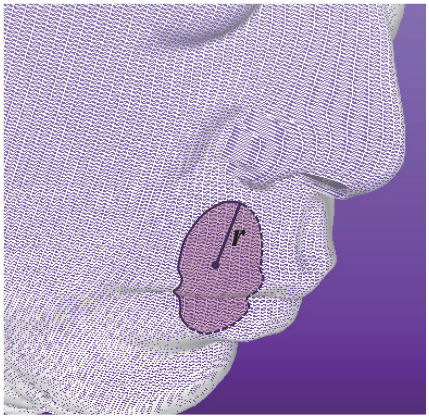


Figure 4.3: The local patch defined on the face model. The local patch is defined as a set of 3D points within the predefined radius r from the centroid point.

grid matrix, one can easily define and extract a local patch by selecting minimum and maximum of row/column discrete index of pixels. However, in the point cloud, a different scheme for local patch is required due to the lack of regularity. In the proposed method, the local patch is defined as a set of points located within a range r around the centroid point. Figure 4.3 illustrates the local patch defined on the face model. The ordinary shape of the extracted local patch from the surface of face is in circular or elliptical shape. As a distance metric to compute the distance from the points to the centroid, the geodesic distance along the facial surface is employed.

Each of extracted point in the local patch contains its absolute 3D coordinate values. In order to describe all patches extracted from different position on the face, the normalization step is introduced. To normalize the local patch, geometric transformation is applied to the points in the local patch. First, they are translated that the centroid of the patch goes to the origin. Then, the points are rotated that the

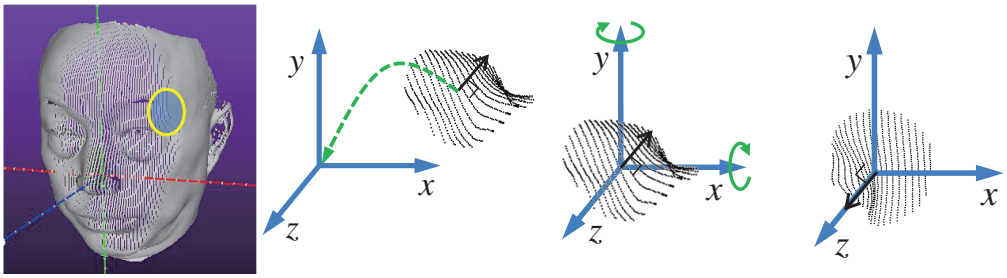
normal direction is equal to the z-axis direction. For the rotation, the normal vector of the patch is calculated by applying the principal components analysis (PCA) to the coordinates of the points near the centroid. The third principal component is the normal direction of the patch. Let us denote the points in a local patch as a matrix x. Each column vector of the matrix denotes each 3D point. The normalized patch $\hat{x}$ is obtained by multiplying the translation $\mathbf{T}$ and rotation matrix $\mathbf{R}$ to the original points x.

$$\hat{x} = \mathbf{N}x = \mathbf{RT}x \tag{4.7}$$

Figure 4.4 illustrates the normalization process of local patch. Here, the normalizing transform $\mathbf{N}$ is also stored with the corresponding patch for the later use. The normalization process has advantage that we can greatly reduce the variability of the patches, since the patches extracted from different centroid coordinate position and direction are stored in the same normalized domain. The underlying idea of the patch normalization process is analogous to the 'Markov assumption over the frequency', which is one of the assumption in the intensity image super resolution methods [16, 48]. They assume high frequency detail is independent of the low frequency value given the mid frequency signal as $P(H|M, L) = P(H|M)$.

4.4.2 Likelihood

The likelihood data cost term $E_d(x_k, y_k)$ is designed to reflect the difference between the low resolution patch and the high resolution patch. The comparison is performed in the aforementioned normalized domain, and the exemplar high resolution patch



(a) (b) (c) (d)

Figure 4.4: Local patch normalization process. (a) A local patch (yellow lined circle) is extracted from the face surface. (b) The translation is applied to all points in the patch so that the centroid of the patch coincide with the origin. (c) The surface normal of the patch is obtained by PCA applied to the points near the centroid. The patch is rotated to make the normal direction coinside with the z-axis. (d) Normalized local patch.

is down-sampled before the data cost calculation. As a data cost, a local surface descriptor [49] is employed. The descriptor takes the 'snapshot' of the target surface with a virtual depth camera located perpendicular to the surface. Although many descriptors for 3D shape are proposed [50, 51, 52, 53], an easy local description is required for the purpose of the proposed method instead of the description of the global structure. One of the advantages of the snapshot is the speed of calculation, which is essential to perform on the large database. The efficiency comes from the simple procedure of description. The scattered 3D points are projected to the virtual image plane and the holes are filled by applying interpolation. In the proposed method, we add minor modifications to the original work [49] for better performance and implementation. First, we implement the camera as affine camera model (camera at infinity) while Malassiotis [49] assumed pinhole projective camera. Second, the natural nearest interpolation [54] is applied instead of linear interpolation method. In the proposed method, central $20mm \times 20mm$ rectangular region of the patch is captured in a 40-by-40 pixels image, and they are concatenated into a 1600-dim descriptor vector. An example of descriptor is displayed in the figure 4.5.

In the proposed method, to measure the snapshot distance between low resolution input patch x_k and the high resolution patch y_k extracted from the database, the y_k should be down-sampled. Considering that they are normalized before applying the descriptor, the snapshot distance between x_k and y_k is formulated as follows:

$$d_{\text{snapshot}}(x_k, y_k) = \|\text{snapshot}(\hat{x}_k) - \text{snapshot}(\hat{y}_k^{\downarrow})\|_2. \tag{4.8}$$

Figure 4.6 shows an example of snapshot descriptor extracted from down-sampled

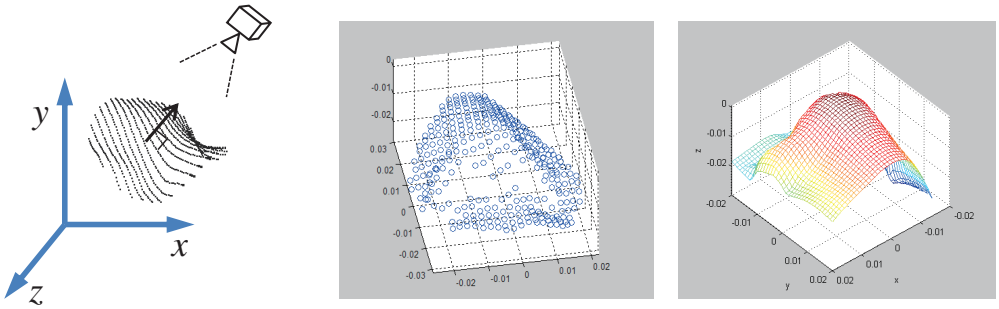
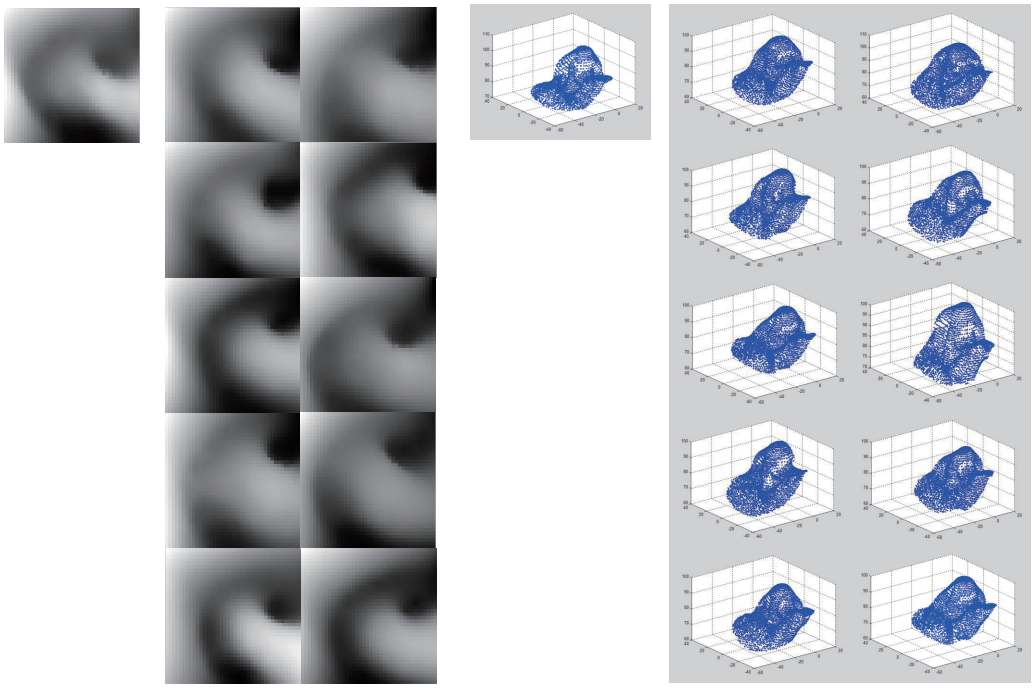


Figure 4.5: Left: Snapshot descriptor describes the local 3D surface by taking a 'snapshot' by a virtual depth camera. Center: 3D points of the local patch to be described. Right: The snapshot descriptor is extracted as a depth image.

patch and the nearest descriptors in the database and their corresponding high-resolution patches.

Although the local shape descriptor has discriminative power to retrieve similar patches in the exemplar database, we can exploit more information about the local patches based on the general shape arrangement of human face. In addition to the snapshot descriptor, the location of the local patch among the global face is considered to calculate the data cost value. Since the shape of human face is generally concave, we interpret the centroid of the local patch in the spherical coordinate system. Our assumption is that the scale variation of human face is not severe, therefore the location is represented only with geometrical azimuth and altitude, dropping the radius value. Thus, to encode the location information into the data cost, the distance between two points is calculated by their angular position in the unit sphere. The shortest geodesic distance between to points on the surface of sphere is the great circle or orthodromic distance, which is a common concept in geographic



(a) (b)

Figure 4.6: (a) Left: An input snapshot decriptor represented as an image. Right: 10 nearest snapshot descriptors in extracted from the database. (b) Left: The true high resolution patch of the input snapshot descriptor. Right: High resolution 3D patches of corresponding snapshot descriptors.

navigation system. Consider two points in the space and let φ_1, θ_1 and φ_2, θ_2 be the azimuth and altitude of two points 1 and 2, respectively[1]. Their absolute difference is given as $\Delta\varphi$ and $\Delta\theta$. Then, the great arc distance is determined only by the central angle between them on the unit sphere surface. The central angle $\Delta\rho$ between two points is calculated by the following formula:

$$\Delta\rho = \arccos\left(\sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2 \cos\Delta\theta\right). \tag{4.9}$$

Since the formula in equation 4.9 can yield numerical rounding error in computer system, a complicated form also has been propose such as:

$$\Delta\rho = 2\arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos\varphi_1 \cos\varphi_2 \sin^2\left(\frac{\Delta\theta}{2}\right)}\right). \tag{4.10}$$

In our implementation, we use simple vector version of the formulation, utilizing the normal vectors and the dot product and cross product of them as follows:

$$\Delta\rho(x, y) = \arccos\left(\mathbf{n}_1 \cdot \mathbf{n}_2\right), \tag{4.11}$$

$$\Delta\rho(x, y) = \arcsin\left(|\mathbf{n}_1 \times \mathbf{n}_2|\right). \tag{4.12}$$

The combination of them is known to be stable for all angles [55], which is employed in our implementation as follows:

$$\Delta\rho(x, y) = \arctan\left(\frac{|\mathbf{n}_1 \times \mathbf{n}_2|}{\mathbf{n}_1 \cdot \mathbf{n}_2}\right), \tag{4.13}$$

where $\mathbf{n}_1$ and $\mathbf{n}_2$ are the vectors of centroid of x and y, respectively.

[1]Note that the azimuth φ is different from the unary potential ϕ.

Assuming that the local patches extracted from similar facial parts are distributed in the Gaussian Parzan window, the $d_{\text{gcd}}(x_k, y_k)$ is formulated with involving the variance parameter as follows:

$$d_{\text{gcd}}(x_k, y_k) = \frac{\Delta\rho(x_k, y_k)^2}{\sigma^2}. \tag{4.14}$$

Summing up, the snapshot descriptor distance which reflects the local shape difference and the great circle distance that measures position disparity between two patches, the final total data cost is formulated as follows:

$$\begin{aligned} E_d(x_k, y_k) &= d_{\text{snapshot}}(x_k, y_k) + d_{\text{gcd}}(x_k, y_k) \\ &= \|\text{snapshot}(\hat{x}_k) - \text{snapshot}(\hat{y}_k^{\downarrow})\|_2 + \frac{\Delta\rho(x_k, y_k)^2}{\sigma^2}. \end{aligned} \tag{4.15}$$

4.4.3 Prior

The prior term enforces neighboring patches to have smoothly varying shape boundary between them. We sample the local patch with a sufficient radius r so that the radius is generally larger than the half of the distance to the neighboring patch's centroid. Of course, to yield a dense output point cloud without holes on the surface, the patch sampling rate should be high enough to provide overlapped patches.

In the overlap region between the neighboring high resolution patch pair y_i and y_j, the proposed algorithm measures the disparity of points to define the smoothness term $E_s(y_i, y_k)$. Since the points in the cloud are not in the regular grid structure, they are not one-to-one matched when they overlap. Thus, we have to guide the points in the overlap region to have matches between the patch pair. The simplest

trick is measuring the closest distance from each point in the current patch to the points in the other patch. However, this idea can generate mistake when a single odd point is located too close or too far from the other patch points. Therefore, we define the smoothness term as the average Euclidean distance between the *mutual nearest neighbor* points in both neighboring patches. Let $O_{i,j}^m$ and $O_{j,i}^m$ be the m-th mutual nearest neighbor point pair in patch y_i and y_j, respectively, the smoothness term is described as follows:

$$E_s(y_i, y_j) = \frac{1}{M} \sum_m \|O_{i,j}^m - O_{j,i}^m\|_2, \tag{4.16}$$

where M denotes the total number of mutual nearest neighbor point pairs.

4.5 Implementation

4.5.1 Training Data

For the training set, we collect high resolution 3D face models captured by laser scanner. The captured data is cleaned and aligned to the canonical direction and position. Local surface patches are sampled from the collected 3D faces and the extracted patches are utilized to build the exemplar database.

4.5.1.1 Pre-Processing

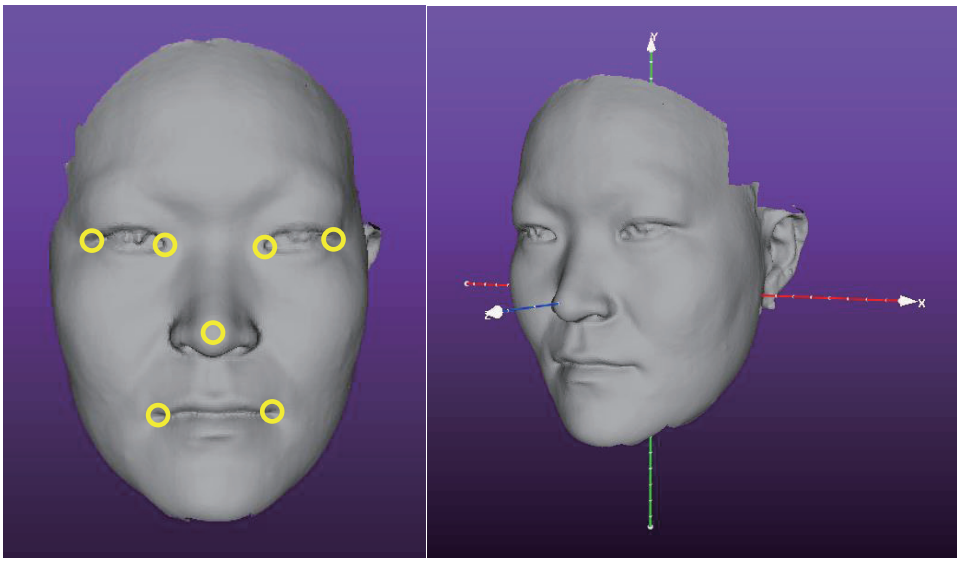
The original scan data is firstly pruned by removing noisy parts such as flying outlier points, unnecessary hair, cloth, and even the scanner itself. Next, the 3D face is aligned to the same orientation. This alignment step is particularly important if we want to utilize the global location information of the local patches. Seven

control points are defined based on the human facial anatomy, { nose tip, left medial canthus(inner corner of an eye), left lateral canthus(outer corner of an eye), right medial canthus, right lateral canthus, left corner of the mouth, right corner of the mouse }. Utilizing the coordinates of the selected control points, the faces are rotated and translated so that all the faces are looking toward the z-axis direction. We do not try any scaling or warping to accurately fit the control points, since the cost function related to great circle distance is loosely designed to cope with small variation. Figure 4.7 displays the control points and aligned face.

4.5.1.2 Patch Extraction

The exemplar database is composed of normalized high resolution local patches $\{\hat{y}\}$ and the corresponding descriptors extracted from the down-sampled patches $\{\text{snapshot}(\hat{y}^{\downarrow})\}$. Also the normalizing transform matrices $\{\mathbf{N} = \mathbf{RT}\}$ and centroid positions of the patches $\{\mathbf{n}\}$ are saved as side information. Given the sampled patch from the original training data y_{train}, the normalization process follows the equation 4.7 as $\hat{y} = \mathbf{RT}y_{\text{train}}$. Our observation is that for the flat areas, i.e. cheek or forehead, the most of the extracted patches do not contain rich shape detail. Storing such non-informative but massive patches does not dramatically increase the performance of the algorithm. Instead, they occupy huge amount of memory space. Thus, we improve the inefficient brute-force sampling, introducing a smarter patch sampling scheme.

Our strategy is sampling more patches where the surface curvature is high, less patches where the surface has low curvature value. The surface curvature is calculated at every point of the training face by the method introduced in [56, 57]. The



(a) Control points (b) Aligned face

Figure 4.7: To align the face models, 7 control points are manually annotated. The yellow circles in (a) depict the position of the defined control points. They are defined on { nose tip, left medial canthus(inner corner of an eye), left lateral canthus(outer corner of an eye), right medial canthus, right lateral canthus, left corner of the mouth, right corner of the mouse }, respectively. The aligned face is displayed in (b).

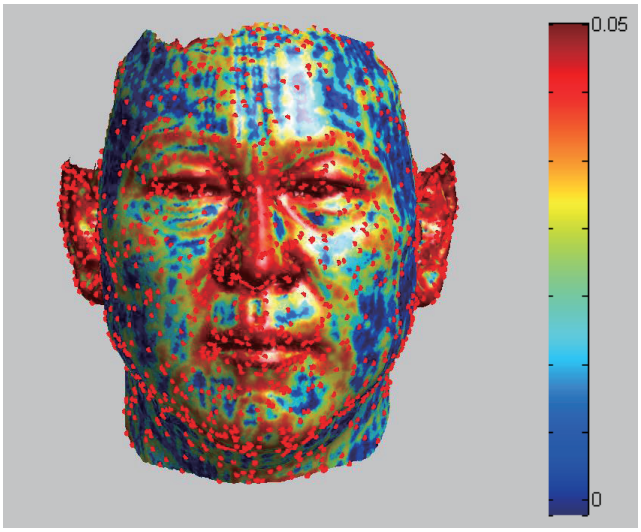


Figure 4.8: Local patch sampling for the exemplar database. The patches are sampled so that their density is proportional to the surface curvature. The color indicates the curvature, and the red dots imply the positions of sampled centroids.

sampling is done probabilistically in the roulette wheel selection scheme. The probability of a point to be selected as a centroid of a patch is proportional to the curvature of the point. As depicted in the figure 4.8, the probabilistic sampling scheme extracts more samples from the high-curvature area than the uniform sampling.

4.5.2 Building Markov Network

The MRF model shown in figure 4.2 illustrates typical regular grid structure for the purpose of explanation. However, in the proposed method, we have to construct proper structure of MRF graph which covers the surface of input 3D face. Since the structure of MRF defines position of local patch as the nodes and the neighborhood relation as the edges between them, efficient graph is essential to yield plausible

super resolution result through the optimization process.

Given the low resolution input 3D face, our strategy is to sample centroid points and the corresponding local patches on the geodesic distance on the surface. First, a random point in the input surface is selected as the first centroid. Next, the geodesic distance from the first centroid point to the all the other points are calculated. The second centroid point is selected as a point which has the farthest geodesic distance, then the selected point is added to the centroid pool. Again, the geodesic distance is calculated from the centroids in the pool to the other points, and select the farthest point as the next centroid point. The procedure is iterated until we have desired number of centroids and corresponding local patches. Following the strategy above, the resulting centroids are pseudo-uniformly distributed along the geodesic surface of input 3D face.

The MRF edges are also constructed utilizing the geodesic distance on the surface. We compute the Voronoi tessellation on the geodesic distance of the extracted local patch centroids. The edges are defined as the segments of the Delaunay triangulation, which is the dual of Voronoi tessellation. Figure 4.9 illustrates the constructed MRF nodes and edges on the input 3D surface by the pseudo-uniform sampling and by computing the Voronoi tessellation.

4.5.3 Reconstructing Super-Resolved 3D Model

If a high resolution patch y is selected as a latent variable candidate of an observed low resolution input x, the original y_{train} and x are located in similar global spherical location, and the down-sampled $y^{\downarrow}$ and x are close in the local descriptor space.

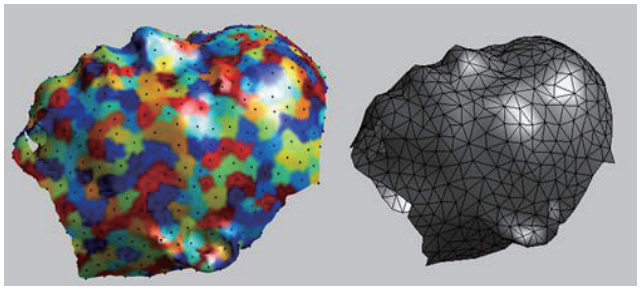


Figure 4.9: Illustration of Markov network defined on the input low resolution 3D facial surface. Left: Selected centroids and their Voronoi cells. Small dots represent centroid points selected by pseudo-uniform sampling. The Voronoi cells are visualized in different colors. Right: Established MRF structure on the face. The triangles are Delaunay triangles which is the dual of Voronoi cells. Note that this figure displays a sparse MRF structure for the purpose of explanation.

Recalling that the local patches are normalized before the snapshot is described, their comparison is performed in the *normalized* domain. Thus, to reconstruct the super-resolved 3D model, the normalized high resolution patch $\hat{y}$ in the database should be projected onto the low resolution patch x by applying inverse of normalizing transform of x. The patch in database is unnormalized as follows:

$$\mathbf{N}_x^{-1}\hat{y} = (\mathbf{R}_x\mathbf{T}_x)^{-1}\hat{y} = \mathbf{T}_x^{-1}\mathbf{R}_x^{-1}\hat{y} = y, \tag{4.17}$$

where $\mathbf{N}_x$ denotes the normalizing transform of x.

When the selected high resolution patches are projected, they will surely have overlap with each other, except the patches on the facial boundary. In conventional intensity image approach, the overlapping pixels are averaged to yield super-resolved final result [16]. In the depth images, Aodha et al. [40] had reported that such

averaging can cause artifacts. In the proposed framework, the overlapping area is difficult to define exactly due to the irregularity of the point cloud, and the dense sampling makes the patches be overlapped heavily. Therefore, we remove all the overlapping points in the final super-resolved point cloud result. As a post-processing to the proposed framework, the points in a patch is simply removed if they are closer to the centroid of the other patches.

4.6 Experiments and Results

For the experiments, total 95 laser scanned faces are procured. The faces are captured from different individuals. The subjects are all Korean in varying age and gender. Among 95 faces, 85 of them are processed to build exemplar, and remaining 10 faces are used as test data. We apply TRW-S as the optimization method [58], which is known to be stable for various optimization problem containing many non-submodular energy terms. Throughout the experiments, the number of candidate is set to 10, and the σ which determines the leverage of global position information is set to 0.2 for the laser scanned data We changed the parameter for the Kinect data as $\sigma = 0.05$, and the snapshot decriptor is applied to the enlarged $30mm \times 30mm$ region.

4.6.1 Quantitative Results

The performance of the proposed method is quantitatively measured by comparing the super-resolved output result with the ground truth high resolution data. To measure the quantitative result, we used the down-sampled 3D face point cloud

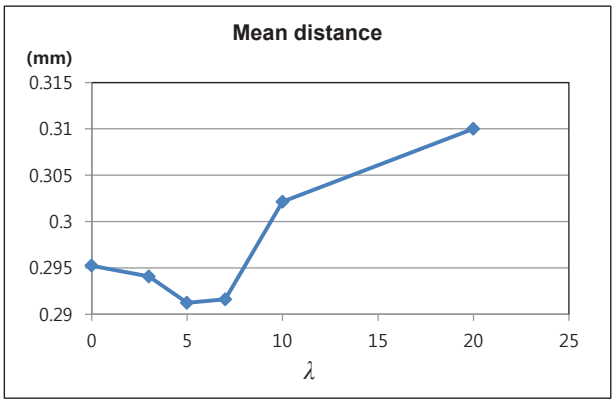


Figure 4.10: Performance on varying λ values. On $\lambda = \{0, 3, 5, 7, 10, 20\}$, the mean distance(error) are 0.295, 0.294, 0.291, 0.292, 0.302, and 0.310, respectively.

model as test data, and the original high resolution version as ground truth. As mentioned earlier, the points in the cloud are distributed irregularly in the space. Therefore, to verify the quality of results, simply measuring the distance between nearest points of both ground truth and super-resolved output can raise unwanted problem. Based on the idea that the ground truth data contains mesh surface, we calculate the distance between the output points and their nearest ground truth surface as the quantitative measure.

In the MAP(Maximum a Posteriori) estimation, the coefficient λ between unary and pairwise costs can affect the quality of solution by controlling the balance of them [59, 60]. To find the optimal value of balancing parameter λ, we also conducted experiments with varying λ. Small subset of test data (3) are selected to observe the trend of output quality with controlled λ. The graph in figure 4.10 summarizes the result. Various $\lambda = \{0, 3, 5, 7, 10, 20\}$ are tested, where $\lambda = 0$ means the

Table 4.1: Super resolution performance from various up-sampling ratio and database size. The units are in *mm* (millimeter).

	Mean distance			RMSE		
	×5	×10	×20	×5	×10	×20
9 DB	0.3182	0.3573	0.3828	0.5663	0.6257	0.6578
20 DB	0.3110	0.3482	0.3715	0.5605	0.6148	0.6463
47 DB	0.2917	0.3325	0.3603	0.5348	0.5995	0.6322
85 DB	0.2900	0.3195	0.3575	0.5420	0.5772	0.6312

Winner-Takes-All (WTA) solution without any consideration of smoothness. This test is conducted in ×10 up-sampling ratio. Following the tendency revealed on this experiment, the value of λ is set to 5 throughout the entire experiments.

To evaluate the performance of the proposed method throughly, the experiments are conducted on 3 different super resolution ratio. We build database for ×5, ×10, and ×20 up-sampling. The effect of database size to the quality of result is also examined. The average distance and the RMSE (Root Mean Squared Error) results for different up-sampling ratio and database size is presented in table 4.1 and figure 4.11. As shown in the table and figure, the low resolution faces are successfully up-sampled with small quantitative error, ranging from 0.29mm to 0.3829mm in average distance and from 0.5348mm to 0.6578mm in RMSE measure. The result also shows that the smaller ratio of up-sampling on the larger database generally tends to grant quantitatively accurate super resolution results.

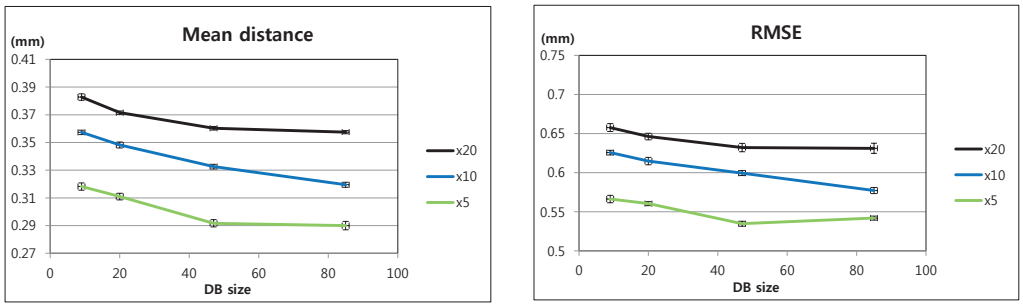


Figure 4.11: Quantitative evaluation results for various up-sampling ratio and database size. The up-sampling ratio of $\times 5$ $\times 10$, and $\times 20$ are performed with exemplar database built from 9, 20, 46 and 85 training faces.

4.6.2 Qualitative Results

The resulting 3D face obtained by applying the proposed super resolution method is visualized in figure 4.12. Three inputs with different down-sampling ratio are used as input data, and they are up-sampled to have the same resolution with original data. As revealed in the figure 4.12, the result is visually more pleasing when the variation of resolution is small. Figure 4.13 shows some close-up images of the result to display the detail of enhanced face model. The figure shows that severe degradation occurs on the facial parts such as nose, mouth, eye and ears of the low resolution 3D model. As depicted in the figure, the facial parts, which have most important curvature and perceptual features, is enhanced by the proposed method.

As a link between the qualitative result and the quantitative measure, we display the errormap in figure 4.14. The distance to the ground truth is calculated per points, and the error level for each 3D point is displayed in the corresponding color. The

shown errormap indicates that large disparity error occurs mostly near the nose and the ears. This result is reasonable, since these curved areas contain richer information compared to the planar areas, and are difficult to estimate the true shape. At the same time these parts are the areas where the biggest improvement occurs when the super resolution algorithm is performed to the degraded low resolution input.

Since the training faces are obtained by the same equipment, we examined the proposed algorithm to the 3D face model captured by another device. We performed the proposed method to the 3D face model obtained by the Artec3D scanner, which is a commercial 3D scanner utilizing the structured light scheme. The input model is down-sampled to have 1/10 number of original points, and up-sampled by the proposed super resolution algorithm. Figure 4.17 shows the resulting super-resolved face of proposed algorithm. As shown in the figure, the information extracted from the training database composed of laser scan data is successfully transferred to the 3D data obtained by other device, implying that the proposed method is not dependent on the characteristics of specific device. In this experiment, the quantitative measure is 0.337 mm of mean distance and 0.6969 mm in RMSE.

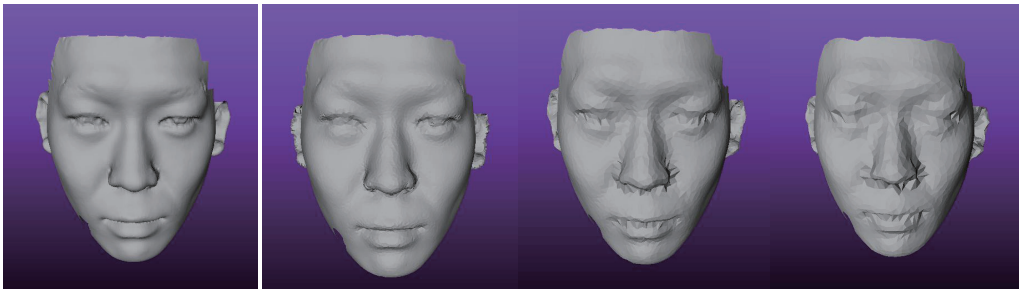
Another very popular 3D imaging device is the Microsoft Kinect sensor which is also based on the structured light system. As the Kinect is originally developed as an input device for the video game, it has very high frame rate and equipped with real time human pose estimator. However, the accuracy of the depth is not accurate in comparison with other devices. To test the proposed method to the 3D model obtained by Kinect, we scanned human face by the Kinect with KinectFusion algorithm [61]. The images in the first row of figure 4.18(a) show typical shape of

human face scanned with Kinect. The Kinect has controllable resolution up to the VGA (640×480), however as we can observe in the figure 4.18 and 4.19, the obtained surface is rather smoothed despite several sweeps of scan with great care. We applied the proposed super resolution method to the 3D models obtained by Kinect, and yielded substantial improvement of the facial detail. In this experiment, the up-sampling ratio is around $\times 2.5$ to $\times 6.5$ according to the scan resolution, however the scanned raw surface has similar level of smoothing, irrespective of the resolution. Figure 4.18 and figure 4.19 show the results of proposed super resolution method applied to the Kinect data. The enhancement in facial parts such as mouth, nose, eyes, and ears is clearly visualized in the zoomed figures. Note that the results can be evaluated only qualitatively since the ground truth is not available.

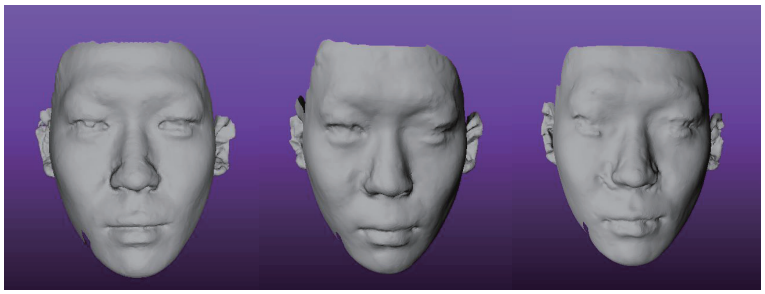
4.7 Summary

The super resolution method for 3D face model represented in point cloud data is proposed in this chapter. The exemplar-based approach is successfully applied to the low resolution single 3D model. The proposed method resolved the super resolution problem by estimating MAP solution of the MRF graph defined on the 3D surface of human face. The proposed super resolution including the local surface descriptor, the global location information, and the curvature based patch sampling over the geodesic distance scheme resolved the particular difficulties coming from the non-regularity of points in the cloud.

The qualitative results and the extensive quantitative evaluation presented the performance of the proposed method, and also showed that the proposed algorithm is

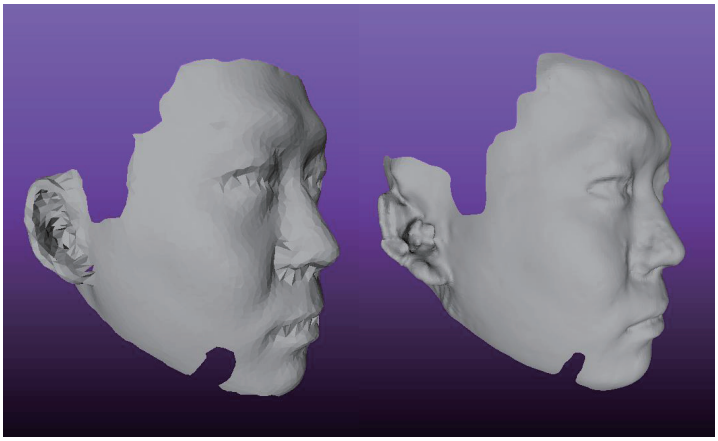


(a) From left to right: Ground truth, Sub-sampled inputs of 1/5, 1/10, and 1/20.

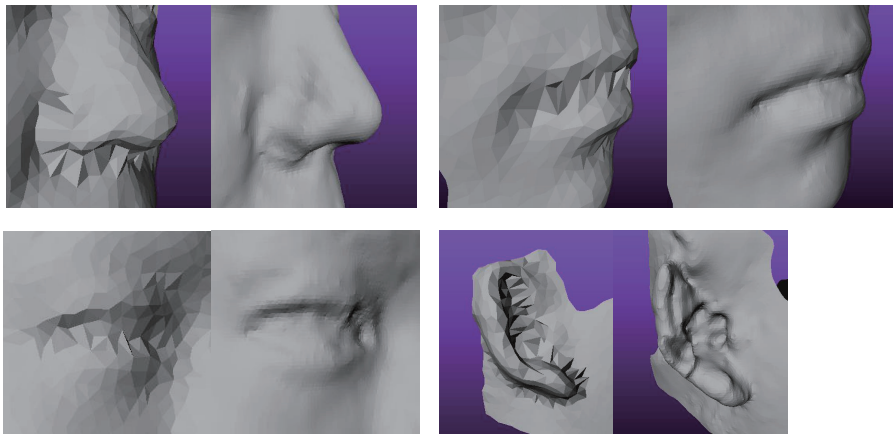


(b) Super resolution results.

Figure 4.12: Super resolution results of $\times 5$ $\times 10$, and $\times 20$ up-sampled faces.

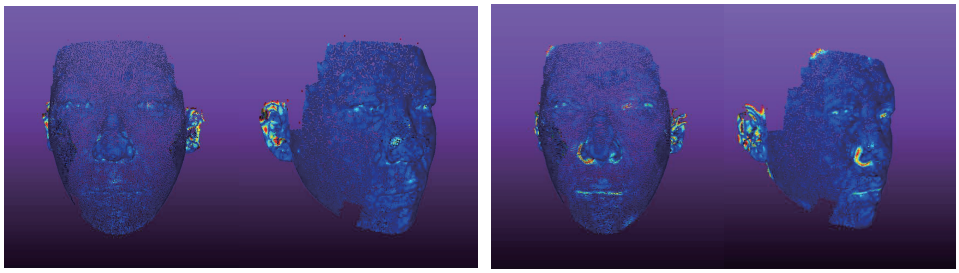


(a) Left: Input. Right: Super resolution output.

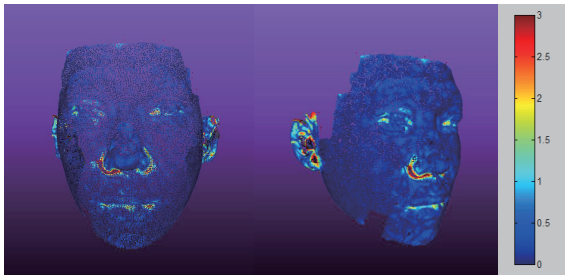


(b) Close-ups. From upper left to lower right, Nose, Mouth, Eye and Ear, respectively.

Figure 4.13: Side-view faces and zoom-up pairs showing the achieved resolution enhancement. The images are captured from $\times 10$ up-sampling results.

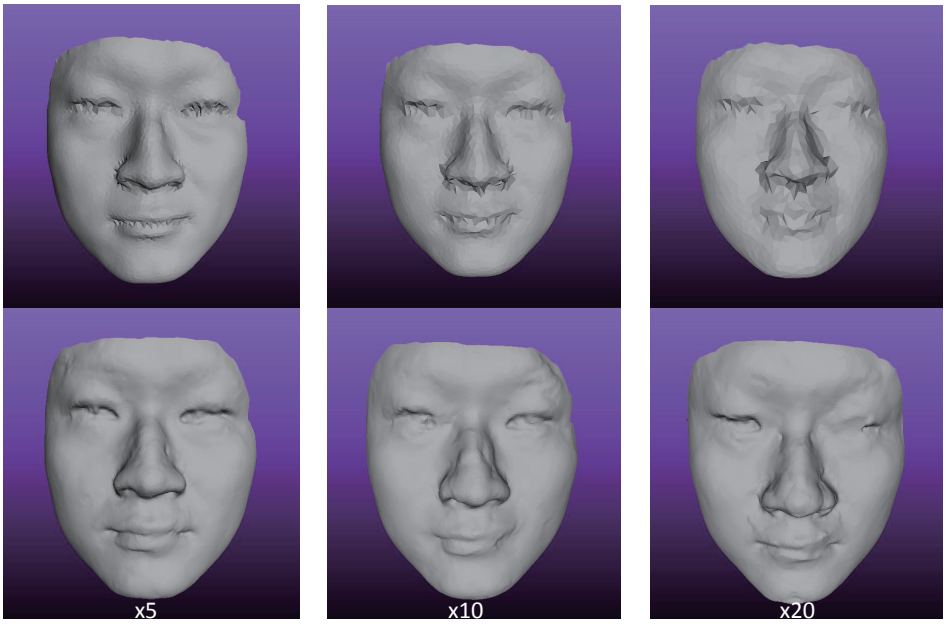


(a) ×5 (b) ×10



(c) ×20

Figure 4.14: Errormap of super-resolved results with specified up-sampling ratio.



(a) Input and super-resolved results.

(b) Original.

Figure 4.15: Qualitative result of a subjective. Tested $\times 5$, $\times 10$, $\times 20$ magnification.

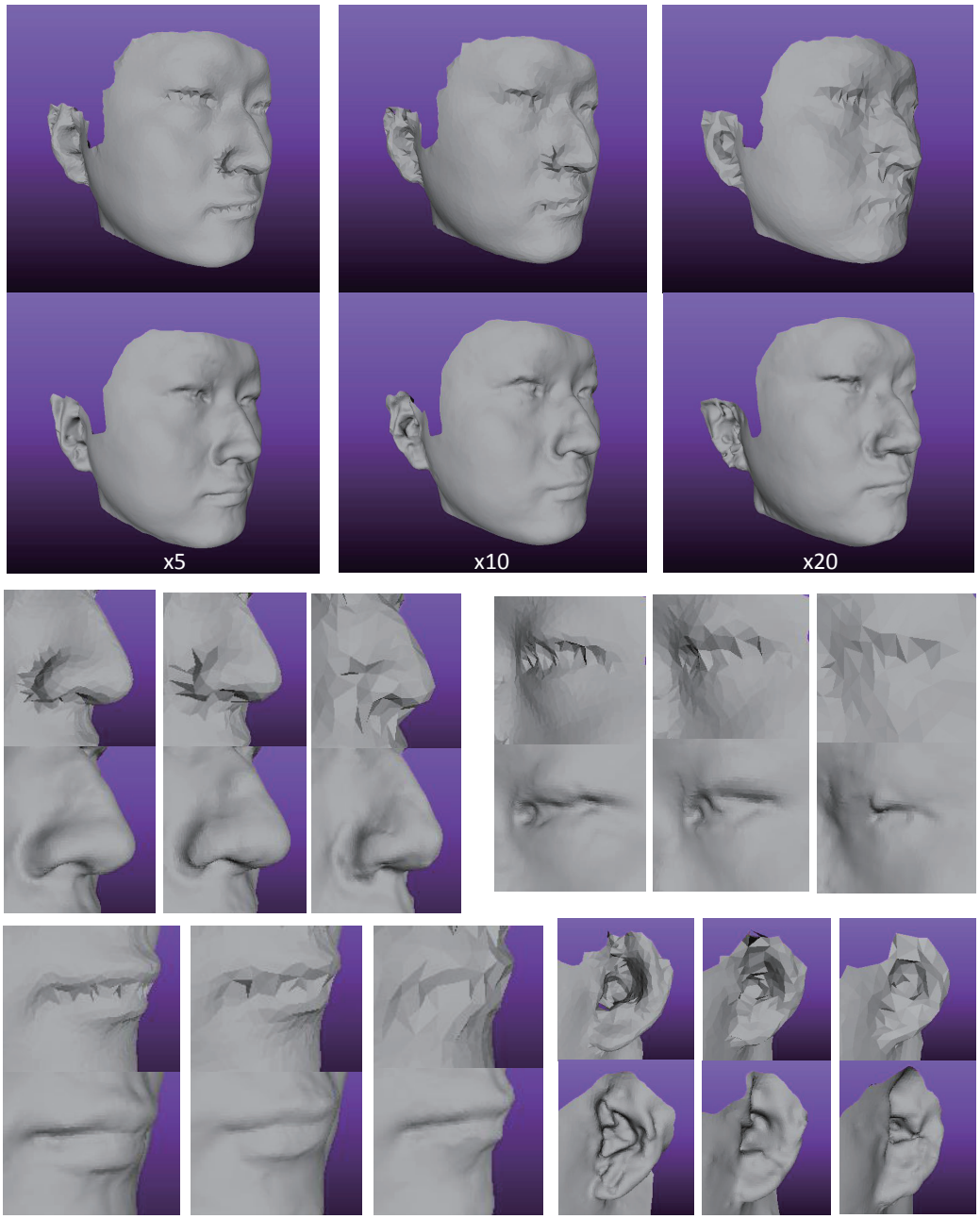
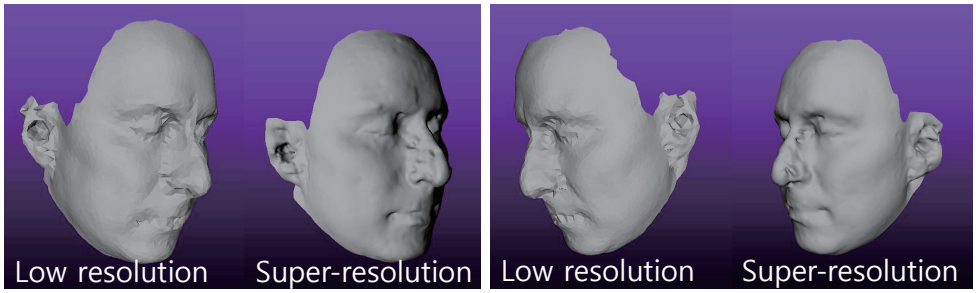
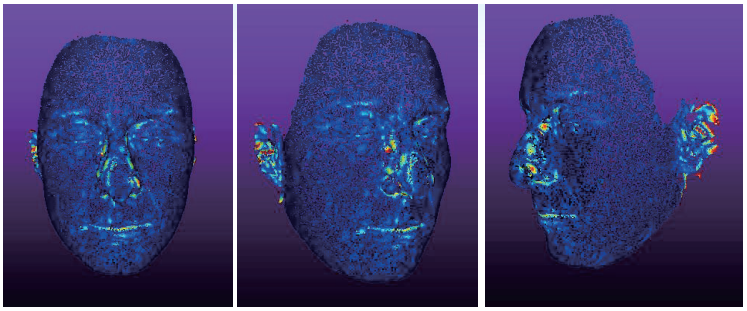


Figure 4.16: Zoomed result of a subjective. Tested $\times 5$, $\times 10$, $\times 20$ magnification.

(a)

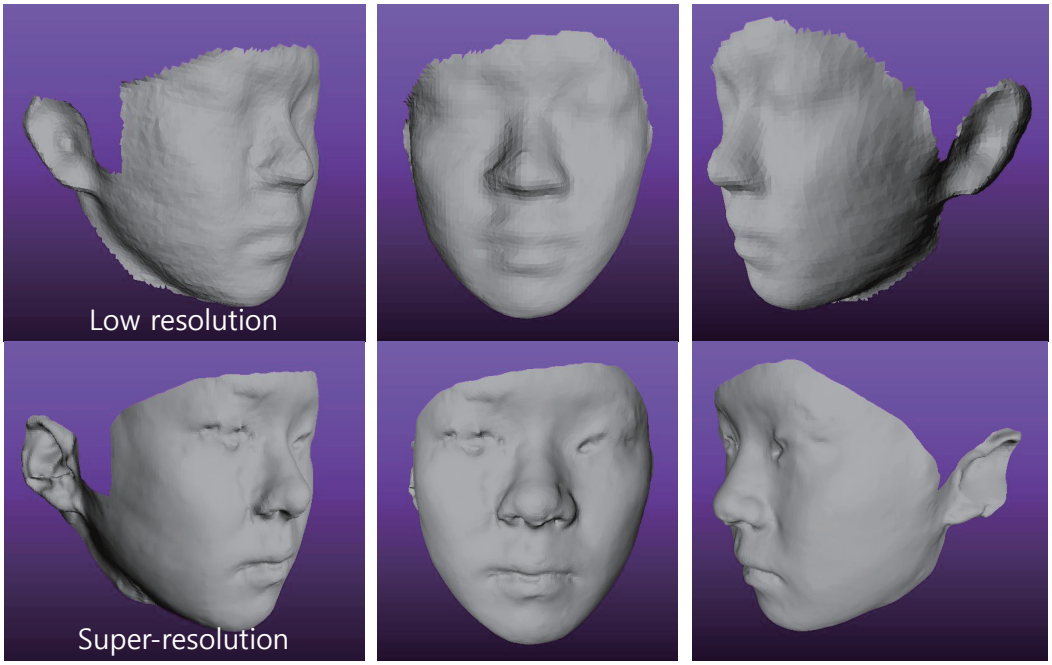


(b)

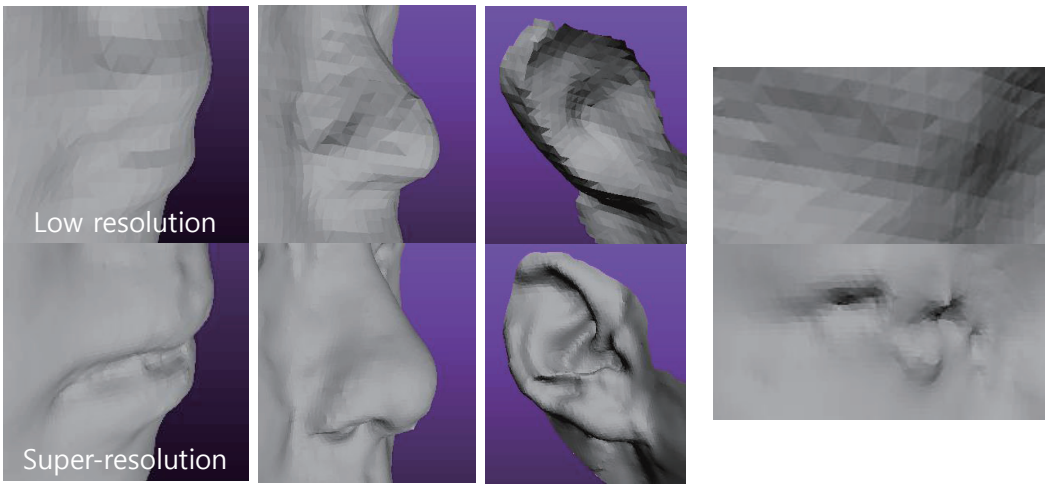


(c)

Figure 4.17: Super resolution result applied to the face model obtained by Artec 3D scanner. The original model is down-sampled by 1/10 points, and up-sampled $\times 10$ by the proposed algorithm. (a), (b) The rendered visualization of front and side views. (c) Errormap of super-resolved result.

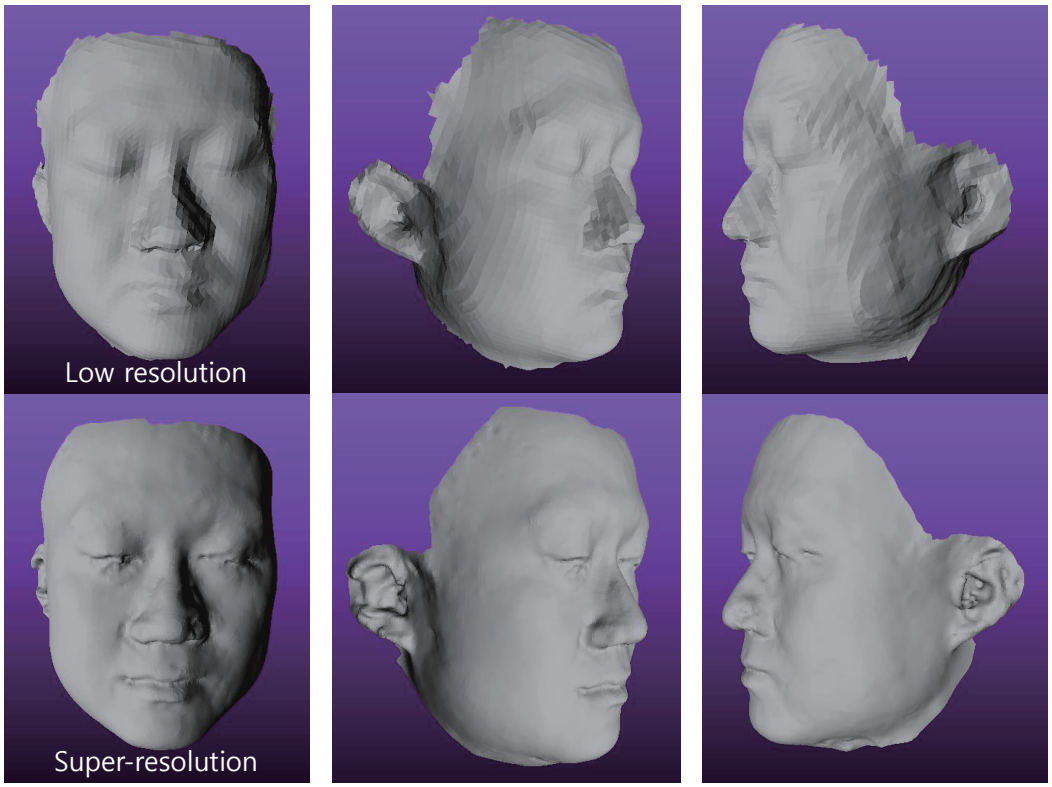


(a)

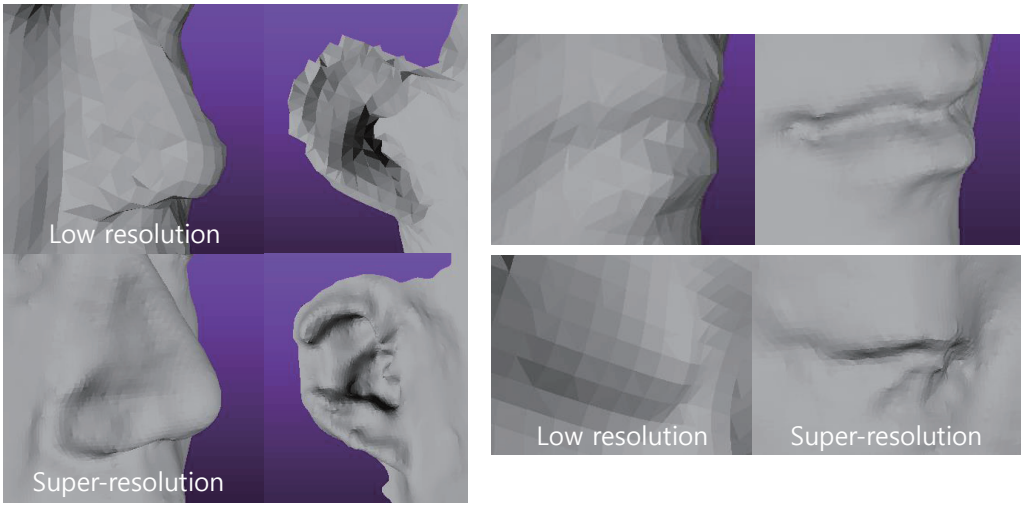


(b)

Figure 4.18: Super resolution of Kinect data. (a) Overall shape of scanned low resolution face and super-resolved result. (b) Close-up figures of mouth, nose, ear, and eye.



(a)



(b)

Figure 4.19: Super resolution of Kinect data. (a) Overall shape of scanned low resolution face and super-resolved result. (b) Close-up of nose, ear, mouth, and eye.

applicable between the training data and test data obtained by the different devices.

Chapter 5

Conclusion

5.1 Summary of Dissertation

In this dissertation, the 3D reconstruction framework for multiple objects in dynamic scenes is presented in chapter 3. In particular, the recognition system based on the object recognition approach is addressed for the handing of abrupt independent motion of objects between images or video frames. Considering that the most of computer vision problem is interpreted as the correspondence matching problem, finding the feature level and object level correspondence is a great advantage to solve the problem. In the proposed reconstruction system, the initial local feature matches containing many false-positives are given as the basis information. As the iterative algorithm explores the given input images without supervision or interaction, the initial matches are pruned and augmented to yield reliable and abundant local matches. At the same time, the clustered local matches provide object-level correspondence and localization information. We bring these information to the 3

dimensional vision problem. The multi-body dynamic scene, which does not satisfy the static constraint, is solved by utilizing the object-level correspondences. Each feature is matched in the multiple view geometry framework only with the features in the same object identity. The obtained 3D calibration information is again utilized to refine the object region segmentation. We turn the interactive seeded segmentation into a non-interactive automatic method by providing reliable seed points which are proven to be stable by the multi-view matching process. The proposed integrated reconstruction framework is applied to the various input data including several still images and videos, where the target objects move continuously or abruptly. The proposed object-centered approach revealed its applicability to the less-controlled dynamic scenes, which were excluded or partially studied topics only with continuous videos.

In chapter 4, a new super resolution method for the 3D face data is presented. In particular, the 3D super resolution method for the point cloud data is addressed. The object-centered point cloud data contains the geometric shape of the target object, irrespective of observing camera position and viewpoint. In the proposed super resolution method, the MAP-MRF is formulated to estimate high-resolution of the 3D face. The low resolution input 3D data is decomposed into local 3D patches, and the Markov network has the local patches as its nodes in the graph. The super resolution is performed by finding the discrete labels for each node in MRF, which correspond to the desired high-resolution patch. To compare and calculate the unary data cost, the local shape descriptor is applied to the local patches of different number of points. We also utilized the global information of the human face

to find better high-resolution candidates in a reasonable search space. The global location is approximately defined by projecting the local patches to the unit sphere, and describing their angular position. The smoothness between neighboring patches is encouraged by the prior term, penalizing the disparity between mutual nearest points of the patch pair. Although the proposed method has analogousness with the conventional intensity image super resolution and the MRF-based low level vision problem, the irregularity of point cloud data is the principal difficulty of applying the traditional method directly to the 3D data. The evaluation of various up-sampling ratio and database is presented, with the result of inter-device information transfer. The application of non-parametric exemplar based approach to the 3D point cloud data super resolution is the main contribution of this study, and to the best of our knowledge, it is the first attempt to the 3D point cloud super resolution.

5.2 Future Works

Although the proposed 3D reconstruction framework in chapter 3 is applicable to various dynamic scenes, it has main assumption in the scenes: the target object should have rigid geometric motion. Segmenting and reconstructing the non-rigid motion is another important research issue. Currently, applying the multiple view geometry technique to reconstruct 3D shape of the non-rigid object requires heavy prior assumptions and has many restrictions [62, 63]. However, as the computer vision technique has been developed to implement the biological visual perception system, the human ability to perceive the non-rigid 3D world only by monocular observation is believed to be possible in computers. In terms of the proposed method

for multiple dynamic scene reconstruction, a more elaborate design of models considering the non-rigidity of the real world objects needs to be investigated. In the super resolution method, although some experiments on the small up-sampling ratio seemed to be saturated in the larger database, we obtained generally improving result as the size of exemplar database grows. The recent research issues about the big data suggest that even the approach is very simple, one can yield superior result with the aid of big data. In the same line of research, developing the proposed algorithm to have the ability to exploit larger exemplar database is meaningful, since utilizing large database without considering the complexity will just return an infeasible problem. Observing the performance on the exemplars from differing age, gender and ethnic group is considerable. Also, if we can retain enough exemplars from the generic objects in the world, and find an efficient way to use the data, the generic 3D super resolution is believed to be a valuable research.

Bibliography

[1] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "Segmenting, modeling, and matching video clips containing multiple moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 477–491, March 2007.

[2] J. Wills, S. Agarwal, and S. Belongie, "What went where," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[3] Y. M. Shin, M. Cho, and K. M. Lee, "Multi-object reconstruction from dynamic scenes: An object-centered approach," *Computer Vision and Image Understanding*, vol. 117, no. 11, pp. 1575–1588, November 2013.

[4] L. Ballan, G. J. Brostow, J. Puwein, and M. Pollefeys, "Unstructured video-based rendering: Interactive exploration of casually captured videos," *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2010)*, pp. 1–11, July 2010.

[5] S. Sinha, J. Kopf, M. Goesele, D. Scharstein, and R. Szeliski, "Image-based rendering for scenes with reflections," *ACM Transactions on Graphics (Proceedings*

of SIGGRAPH 2012), vol. 31, no. 4, pp. 100:1–100:10, August 2012.

[6] J. Wang and E. H. Adelson, "Layered representation for motion analysis," in *Proceedings of IEEE Computer Vision and Pattern Recognition*, 1993.

[7] S. Ayer and H. S. Sawhney, "Layered representation of motion video using robust maximum-likelihood estimation of mixture models and mdl encoding," in *Proceedings of IEEE International Conference on Computer Vision*, 1995.

[8] J. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 625–638, September 1994.

[9] J. Xiao and M. Shah, "Accurate motion layer segmentation and matting," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[10] J. Costeira and T. Kanade, "A multi-body factorization method for motion analysis," in *Proceedings of IEEE International Conference on Computer Vision*, 1995.

[11] R. Vidal and S. Sastry, "Optimal segmentation of dynamic scenes from two perspective views," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[12] Q. Ke and T. Kanade, "A subspace approach to layer extraction," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[13] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical Models and Image Processing*, vol. 53, no. 3, pp. 231–239, May 1991.

[14] L. C. Pickup, S. J. Roberts, and A. Zisserman, "Optimizing and learning for super-resolution," in *Proceedings of British Machine Vision Conference*, 2006.

[15] S. Farsiu and M. D. Robinson, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, October 2004.

[16] W. T., E. C. Pasztor, and O. T. Carmichael, "Learning low level vision," *International Journal on Computer Vision*, vol. 40, no. 1, pp. 25–47, October 2000.

[17] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, March 2002.

[18] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proceedings of IEEE International Conference on Computer Vision*, 2009.

[19] M. Zontak and M. Irani, "Internal statistics of a single natural image," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[20] J. Yang, J. Wright, , T. Huang, , and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 1861–2873, November 2010.

[21] N. Thakoor, J. Gao, and V. Devarajan, "Multibody structure-and-motion segmentation by branch-and-bound model selection," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1393–1402, June 2010.

[22] K. E. Ozden, K. Schindler, and L. van Gool, "Simultaneous segmentation and 3d reconstruction of monocular image sequences," in *Proceedings of IEEE International Conference on Computer Vision*, 2007.

[23] A. W. Fitzgibbon and A. Zisserman, "Multibody structure and motion: 3-D reconstruction of independently moving objects," in *Proceedings of European Conference on Computer Vision*, 2000.

[24] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of International Joint Conference on Artificial Intelligence*, 1981.

[25] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proceedings of European Conference on Computer Vision*, 2010.

[26] W.-C. Lu, Y.-C. F. Wang, and C.-S. Chen, "Learning dense optical-flow trajectory patterns for video object extraction," in *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010.

[27] K. Ozden, K. Cornelis, L. V. Eycken, and L. V. Gool, "Reconstructing 3D trajectories of independently moving objects using generic constraints," *Computer Vision and Image Understanding*, vol. 96, no. 3, pp. 453–471, 2004.

[28] M. Cho, Y. M. Shin, and K. M. Lee, "Unsupervised detection and segmentation of identical objects," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[29] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proceedings of British Machine Vision Conference*, 2002.

[30] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proceedings of European Conference on Computer Vision*, 2002.

[31] F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets, or 'How do I organize my holiday snaps?'," in *Proceedings of European Conference on Computer Vision*, 2002.

[32] M. Brown and D. Lowe, "Unsupervised 3D object recognition and reconstruction in unordered datasets," in *Proceedings of International Conference on 3-D Digital Imaging and Modeling*, 2005.

[33] M. Vergauwen and L. V. Gool, "Web-based 3D reconstruction service," *Machine Vision and Applications*, vol. 17, no. 6, pp. 411–426, December 2006.

[34] D. Martinec and T. Pajdla, "Robust rotation and translation estimation in multi-view reconstruction," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[35] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1998.

[36] T. H. Kim, K. M. Lee, and S. U. Lee, "Generative image segmentation using random walks with restart," in *Proceedings of European Conference on Computer Vision*, 2008.

[37] H. Edelsbrunner, D. G. Kirkpatrick, and R. Seidel, "On the shape of a set of points in the plane," *IEEE Transactions on Information Theory*, vol. 29, no. 4, pp. 551–559, July 1983.

[38] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, August 2010.

[39] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *International Journal of Computer Vision*, vol. 80, no. 2, pp. 189–210, November 2008.

[40] O. M. Aodha, N. D. F. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proceedings of European Conference on Computer Vision*, 2012.

[41] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "Lidarboost: Depth superresolution for ToF 3D shape scanning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[42] A. Rajagopalan, A. Bhavsar, F. Wallhoff, and G. Rigoll, "Resolution enhancement of PMD range maps," *Lecture Notes in Computer Science*, vol. 5096, pp. 304–313, 2008.

[43] Q. Yang, R. Yang, J. Davis, and D. Nist, "Spatial-depth super resolution for range images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[44] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proceedings of Advances in Neural Information Processing System*, 2005.

[45] M. Kiechle, S. Hawe, and M. Kleinsteuber, "A joint intensity and depth co-sparse analysis model for depth map super-resolution," in *Proceedings of IEEE International Conference on Computer Vision*, 2013.

[46] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proceedings of IEEE International Conference on Computer Vision*, 2013.

[47] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, November 1984.

[48] M. R. Luettgen, W. C. Karl, and A. S. Willsky, "Efficient multiscale regularization with application to the computation of optical flow," *IEEE Transactions on Image Processing*, vol. 3, no. 1, pp. 41–64, January 1994.

[49] S. Malassiotis and M. G. Strintzis, "Snapshots: A novel local surface descriptor and matching algorithm for robust 3D surface alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 7, pp. 1285–1290, July 2007.

[50] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 443–449, May 1999.

[51] A. M. Yamany and A. A. Farag, "Free-form surface registration using surface signatures," in *Proceedings of IEEE International Conference on Computer Vision*, 1999.

[52] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik, "Recognizing objects in range data using regional point descriptors," in *Proceedings of European Conference on Computer Vision*, 2004.

[53] D. Zhang and M. Hebert, "Harmonic shape images: A representation for 3D free-form surface based on energy minimization," in *Proceedings of International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 1999.

[54] R. Sibson, "A brief description of natural neighbor interpolation," *Interpreting Multivariate Data*, pp. 21–36, 1981.

[55] K. Gade, "A non-singular horizontal position representation," *The Journal of Navigation*, vol. 63, no. 3, pp. 395–417, July 2010.

[56] D. Cohen-Steiner and J.-M. Morvan, "Restricted delaunay triangulations and normal cycle," in *Proceedings of 19th Annual ACM Symposium on Computational Geometry*, 2003.

[57] P. Alliez, D. Cohen-Steiner, B. L. Olivier Devillers, and M. Desbrun, "Anisotropic polygonal remeshing," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 485–493, July 2003.

[58] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1568–1583, October 2006.

[59] L. Zhang and S. M. Seitz, "Estimating optimal parameters for MRF stereo from a single image pair," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 331–342, February 2007.

[60] U. Schmidt, Q. Gao, and S. Roth, "A generative perspective on MRFs in low-level vision," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[61] R. A. Newcombe, A. J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, 2011.

[62] L. Torresani, A. Hertzmann, and C. Bregler, "Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 878–892, May 2008.

[63] J. Xiao, J. Chai, and T. Kanade, "A closed-form solution to non-rigid shape and motion recovery," *International Journal of Computer Vision*, vol. 67, no. 2, pp. 233–246, April 2006.

국문 초록

본 논문에서는 복수의 물체가 등장하는 동적인 장면의 3차원 형상을 복원하는 기법과, 얻어진 3차원 모델, 특히 인간의 얼굴을 대상으로 해상도를 높이는 기법에 대해 다룬다. 다시점의 영상을 이용하는 일반적인 3차원 복원 기법들은 정적인 장면을 대상으로 적용되는 기법으로서 물체들의 배치는 다시점의 영상을 얻는 동안 고정되어 있어야 한다. 제안하는 기법의 주된 목표는 시점마다 물체들의 배치가 변화하는 보다 일반적인 조건에서 복수 물체의 3차원 모델을 복원하는 데 있다. 제안하는 기법은 비감독적인 상호 인식 기법을 적용하여 동적 장면을 객체 중심으로 분할함으로서 동적 장면에서 기하학적 일관성이 결여되는 현상을 극복한다. 인접하는 프레임 간에는 연속적인 작은 움직임만이 존재한다고 가정하는 종래의 동작 분할 알고리즘과는 달리 상호 인식 기법은 순서 없이 배열되어 넓은 베이스라인을 가지는 시점들로부터 촬영한 영상에서 동일한 물체끼리의 정합을 신뢰성있고 정확하게 찾아낸다. 각각 물체 영역의 경계선을 명확히 분할하기 위해 움직임 기반 구조 복원(Structure-from-motion) 과정에서 얻어진 3차원 희소점들을 이용한다. 3차원 희소점들은 기하학적 관계와 광학적 일관성을 동시에 고려하여 생성되므로 높은 신뢰도를 갖추고 있다. 이 3차원 희소점은 영상 분할 알고리즘의 씨앗점으로 이용되어 대화식 영상 분할 알고리듬을 사용자의 간섭 없이 동작하도록 해 준다. 다양한 영상에서 수행된 실험을 통해 제안하는 기법의 효과성을 보였으며, 특히 기존의 기법과 비교하여 물체들이 갑작스런 움직임을 가지고 있는 영상에서 강인하게

동작함을 보였다.

고해상도의 3차원 모델을 얻는것 역시 중요한 주제이다. 장면의 3차원 복원에 이용된 다시점 영상 혹은 TOF 카메라나 레이저 스캐너 등의 3차원 영상 장비들이 가지고 있는 해상력의 한계로 인해 3차원 자료의 해상도를 높이기 위한 초해상도 기법이 요구되는 실정이다. 본 논문에서는 3차원의 포인트 클라우드(point cloud)로 표현된 단일 얼굴 모델에 대한 초해상도 기법을 다룬다. 카메라 중심의 깊이 영상에 비해 포인트 클라우드는 객체 중심으로 표현된 3차원 데이터라 할 수 있다. 밝기 영상 및 깊이 영상에 적용된 기존의 초해상도 기법에 비해 본 논문은 3차원 포인트 클라우드 데이터에 대하여 초해상도 기법을 적용하였다. 이 과정에서 동일한 물체에 대한 추가적인 밝기 혹은 깊이 영상 관찰 없이 단일한 3차원 포인트 클라우드만을 입력으로 하여 비모수적인 기계학습 기법을 적용하였다. 이 문제는 고해상도-저해상도 짝으로 이루어진 3차원 정보를 이용하여 사전에 학습된 데이터베이스를 이용함으로 해결한다. 3차원 점들 위에서 형성된 마르코프 랜덤 필드의 각 노드의 레이블에 따른 에너지 함수를 정의하고 이를 최적화하는 과정을 통해 초해상도 3차원 모델을 얻는다. 실험을 통해 제안하는 기법이 높은 정확도로 3차원 초해상도 문제를 해결함을 보였다.

Key words: 컴퓨터 비전, 3차원 복원, 동적인 장면, 상호 인식, 다중 물체, 초해상도, 포인트 클라우드

Student number: 2009-30195